A Bioinformatic Study of the 5' Upstream Region of the Human Gene

Dana Cohen
2008

Cancer Research UK
44 Lincoln's Inn Fields
London WC2A 3PX

and

Department of Biochemistry
and Molecular Biology
University College London
Gower Street London
WC2E 6BT

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

I confirm that the work presented in this thesis is my own.
Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Dana Cohen

Abstract

Understanding transcription regulation is a key issue in biology. There is much non-coding sequence upstream of the human gene, a small proportion of which is regulatory. The aim of this project is to gain a better understanding of organisation, structure and function of this region via large-scale sequence studies.

The 10Kb 5' upstream sequence has been analysed for changes at different positional sections along this stretch. There are four main categories of investigation:

1. Dinucleotide composition and representation.
2. Distance from randomness by comparing real and randomised sequences.
3. Sequence similarity using a pattern analysis.
4. Distribution and representation of regulatory motifs.

DNA generally avoids flexible dinucleotide steps. In the upstream sequence the DNA becomes even less flexible towards the start site of transcription. In contrast there is an increase in bistable steps in this direction. It is concluded that these structural changes including enhanced stiff and bistable steps are likely important in transcription regulatory regions.

The weak/strong (W/S) and purine/pyrimidine (R/Y) properties in the upstream sequence are different to each other and even opposing at times. For instance, the R/Y sequence becomes more distant from the random model towards the start site whereas the W/S becomes closer to it. Opposing sequence similarity trends are observed across the upstream sequence for R/Y and W/S. Also, the regulatory motif distribution and representation are very different depending on whether the sequence is viewed as R/Y or alternatively as W/S nucleotides. This is likely due to the different roles of these nucleotide properties within the upstream DNA and more specifically within regulatory regions.

These results may have important implications for the process of direct and indirect readout in protein-DNA binding. It is suggested that the R/Y sequence generally has a greater influence over indirect readout whereas the W/S sequence has more impact on direct readout. Furthermore it is proposed that avoidance of inappropriate (or promiscuous) regulatory protein binding to DNA occurs primarily via the R/Y sequence of the regulatory elements, i.e. via avoidance of indirect readout and the docking step of protein-DNA binding.

Therefore this study of DNA sequence has revealed changes in specific properties across the upstream region of the human gene from which have been drawn conclusions about its role in transcription regulation.

Contents

List of Figures

List of Tables

1. General Background

Almost every cell in the human body contains a complete genome. Yet the cells of the eyes are very different to skin cells and muscles cells etc. How is it possible that the same set of genes give rise to different cell types? The answer to this lies with gene regulation. The human being starts off as a single cell. During development of this cell into the fully-grown human, certain genes are switched on and a subset is switched off. This process is finely tuned with levels of the protein products being carefully regulated in each cell.

The human genome project has provided a large amount of DNA sequence data. This data can be analysed using computational techniques. Only an estimated 1.1% of human genomic DNA is sequence that is spanned by exons (Venter et al, 2001). Therefore only a minority of genomic sequence actually codes for protein in the human. Since coding sequences represent the functional units of the genome which code for the building blocks of the cell, they have been the most widely studied via computational techniques. Regulation of gene expression is crucial to cell function; therefore the regions of the genome responsible for this regulation are also of great interest and importance.

The regulatory regions make up a proportion of the non-coding sequence of the gene and may be found therein. Bioinformatic techniques are extremely useful for gaining an understanding of these sequences since these methods can help derive biological meaning from this data. Understanding gene regulation and what causes the cell to choose to switch on a specific subset of genes at a particular time is essential to understanding cell biology, growth and development.

Regulation of gene expression is the means by which the cell controls the proteins that it makes. This includes the specific expression of a particular mRNA and the quantity/rate at which that protein is generated. The control or regulation may be generally divided into transcriptional and post-transcriptional. For the majority of genes transcriptional control is the most important since this ensures that excess mRNAs are not synthesised. Only transcriptional regulation will be discussed further.

1.1 Transcription Regulation

1.1.1 The 5' upstream regulatory region: Its structure and function

Particular locations of the 5' upstream region of a protein-coding gene are usually responsible for transcription regulation. The 5' upstream region contains the promoter (see figure 1.1) and also other regulatory sequences. The promoter is the most important determinant of transcription regulation. The upstream region and its overall structure and function will be described in terms of three general regions; the core promoter, proximal promoter and the enhancer (or repressor).

The core promoter is the location at which the basal apparatus for transcription assembles. Transcription is initiated at this location together with the RNA polymerase II (PolII) enzyme that synthesises mRNA. The proximal promoter contains binding sites for other (non-basal) transcriptions factors. These may be activators or repressors of transcription. Together the core promoter and proximal promoter form the region that is commonly referred to as the promoter. The enhancer possesses binding sites for transcription factors that act to enhance transcription. This element is typically further upstream than the proximal promoter and is therefore often physically distinct from it. These three regulatory regions their function, cross interactions and boundaries will be discussed in more detail in the following sections.

In general, the upstream regulatory DNA possesses information in the form of signals that are integrated and the sum of which produce an output. These signals constitute protein binding motifs of the DNA. The information is 'read' by the cell via proteins/transcription factors that bind to the DNA. This is true for the core promoter, proximal promoter and also the enhancer/repressor. The sum (or net effect) of these signals both within and between these regulatory elements in the upstream region of a gene determine the nature of the output, namely the mRNA product.

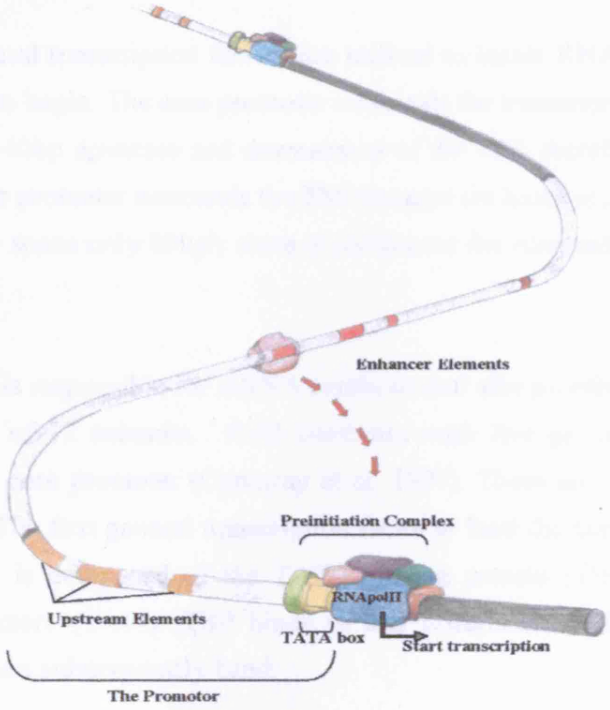


Figure 1.1: Schematic diagram of transcription regulation and the general structure of the upstream elements responsible.
In many cases the TATA box locates the start region. The TATA box binding protein and PolII bind the DNA followed by other basal transcription factors that make up the preinitiation complex. The TATA box combined with other upstream regulatory elements form the promoter. When bound by their transcription factors they either induce or inhibit transcription. An enhancer may be present further upstream that enhances the level of transcript formed. The enhancer is thought to function via factors that bind to it and then looping of the DNA causes interactions between these factors and the initiation complex. The looping of the DNA is indicated in the diagram by the red arrows.

1.1.2 The core promoter

The core promoter is defined as the minimal region of DNA that is capable of directing activator-independent, low-level transcription by PolII. Structural genes require several general (or basal) transcription factors for initiation of transcription (Reese, 2003). These together form the preinitiation complex (PIC) with Pol II. Some transcription factors bind directly to DNA in

the promoter, whilst others bind to already bound proteins within the complex. The basal elements alone though are only capable of promoting initiation of transcription at a low level.

Basal transcription factors are utilised to locate RNA polymerase II in position for transcription to begin. The core promoter surrounds the transcription start site (TSS) and usually spans around 40bp upstream and downstream of the TSS, therefore it typically spans 80bp's in total. The core promoter surrounds the TSS because the basal apparatus actually locates the TSS and it usually spans only 80bp's since it constitutes the minimal machinery required to initiate transcription.

PolII is responsible for mRNA synthesis and also proofreading the nascent transcript. It is comprised of 12 subunits. PolII combines with five general transcription factors which recognise the core promoter (Conaway et el, 1997). These are TFIIB, TFIID, TFIIE, TFIIF and TFIIH. The first general transcription factor to bind the core promoter is TFIID. This is a complex that is composed of the TATA-binding protein (TBP) in addition to some TBP-associated factors (TAFs). TBP binds to and distorts the DNA so that other proteins or complexes can subsequently bind.

The entire machinery composed of all these components is a giant complex of about 60 proteins and a mass of 3 million Daltons. It forms the core of the transcription machinery at the core promoter and has the following functions; 1. Unwinding the DNA, 2. synthesising RNA, and 3. Rewinding the DNA.

Core promoter motifs on the DNA to which the general transcription factors bind are not universal in that they are not located in every promoter within the genome. The core promoter may include the following elements; the TATA box, Inr, DPE, BRE, MTE. The majority of these elements interact with the transcription factor TFIID, although the interaction is in some cases with different subunits of the machinery. These elements will now be described.

TATA-box promoters

The TATA-box is a core promoter element that initiates transcription. Its consensus sequence is TATAWAAR and is usually found 25-30bp upstream of the TSS in eukaryotes. The TATA-box is bound by TBP, the TATA-binding protein subunit of the TFIID complex (Burley et al, 1996), although other proteins may also recognise the TATA-box. TBP though

seems to be the basal transcription factor that often identifies the start site of transcription. When TFIID binds the TATA box it nucleates the PIC formation that includes RNA Pol II.

A wide range of AT-rich sequences can actually function as a TATA-box. This is because TBP is likely to recognise the helix via sequence independent mechanisms (Kim et al, 1993). Whilst the TATA-box is a relatively conserved motif of the core promoter elements, the consensus sequence is present only in an estimated 12% of human promoters (Bajic et al, 2004). This begs the obvious question; how is transcription initiated in TATA-less promoters? In some cases the initiator element Inr provides the answer.

The Inr region

Inr (YYANWYY) surrounds the TSS. It is recognised by the TAF1 and TAF2 subunits of TFDII. Inr can initiate transcription in promoters lacking a TATA box (Smale et al, 1990 and Kaufmann et al, 1994). Although it may also be present in some promoters that do contain a TATA-box, in which case they act in a combined fashion. Inr is found in 55% of human promoters.

Inr can identify the precise location of a TSS. Also, when present in the core promoter together with a TATA-box it assists and determines promoter strength. In other cases its functionality is analogous to that of a TATA box. Therefore a TATA or Inr motif is sufficient for core promoter activity.

Within synthetic promoters Inr can operate either together with the TATA box or with Sp1 activation binding sites that are located upstream of the Inr. Also when Inr is inserted downstream of a TATA box, the transcript level is several times higher than without it. This shows that Inr is capable of either initiating transcription and/or of elevating it. When Inr is relocated either nearer or closer to Sp1, the TSS is also relocated together with the Inr.

Both Inr and to an extent the TATA box have fairly loose consensus sequences. These motifs are frequently present within the genome. The question is; how can transcription initiation rely on these loose consensus sequences? One possible explanation is that location within the core promoter is essential for function. Furthermore, chromatin structure and the existence of other nearby elements on the DNA may be important for identifying these motifs.

Since core promoters may contain either or both the TATA box and Inr, or alternatively neither of these elements, it is possible to conclude that different elements may be utilised for

transcription initiation and core activity. In some cases TATA box and Inr seem interchangeable. What then is the purpose of initiating transcription and core promoter activity with different elements that may also be present in varying combinations?

The answer to this may be revealed to an extent by transcriptional repressors and activators. For instance, p53 and topiosomerase1 repress transcription from promoters with the TATA box but not from promoters with Inr (Aso et al, 1994). Also, activation domains, Sp1 stimulate Inr (but not TATA) promoters. Therefore it may be that the presence of different elements at the core promoter results in either resistance to particular repressors or to subjection to certain activators.

It may be useful to group genes according to the type of element present in their core promoter and to build networks according to resistance or subjection to specific repressors/activators. Although no clear functional connection has yet been discovered for TATA-box or Inr containing promoters, their interaction with and influence by the repressors/activators may provide important clues. This type of analysis would lead to an improved understanding of core promoter function.

Other core promoter elements

The downstream promoter element (DPE; consensus RGWYV) also often plays a role in transcription initiation in TATA-less promoters (Kadonaga et al, 2002). DPE is also recognised by TFDII, but unlike the other elements it is bound by the TAF1 and TAF2 subunits of the protein. DPE usually functions together with Inr (Burke et al, 1996).

The motif ten element (MTE, consensus: CSARCSSAACGS) is found downstream of the TSS. It also functions together with Inr. It may function either in TATA-less or TATA-box containing promoters (Lim et al, 2004). BRE (consensus: SSRCGCC) occurs just upstream of the TATA-box and binds TFIIB. This element can either increase or repress transcription (Evans et al, 2001).

Finally, CpG islands also play a role in promoter activity and are found in about 80% of human promoters. These promoters often lack a TATA-box and the mechanism of their activity is not understood (Blake et al, 1990). CpG island promoters are often associated with housekeeping genes (Carnici et al, 2006). A study has been carried out to identify CpG's in human promoters (Saxonov et al, 2006). This revealed that 72% of promoters were high in CpG and the remainder had a CpG content that is similar to that found genome-wide. More

specifically this high level of CpG was located at the core promoter region. The difference between genes containing high level CpG core promoters and those with normal level CpG's is unknown. One possible explanation is that high CpG promoters are subject to regulation via methylation of CpG's, which is discussed later in this introduction.

Each of the elements that have been described above may be found in some promoters but none occurs in all promoters. Instead the elements are present at varying combinations within the entire set of human promoters. This means that core promoter transcription regulation is diverse within the species. Core promoter activity can be influenced both by proximal promoter elements and also by enhancers. For example, the presence of either a TATA-box or DPE element in the core promoter can influence the interaction of that promoter with enhancers (Butler et al, 2001). These interactions may also be affected or induced by other promoter elements. The sum of these interactions may either activate or repress transcription and are likely to determine the rate of transcript synthesis.

1.1.3 The proximal promoter

The proximal promoter extends upstream of the core promoter (and the initiation complex), usually up to around 300bp upstream of the TSS, however it may extend further, up to 2Kb upstream of the TSS. Transcription factors at the proximal promoter bind to specific sequence elements and it is thought that these transcription factors bind any DNA that contains their target sequence. The proximal promoter contains regulatory motifs utilised by tissue-specific transcription factors. The upstream transcription factors bind to specific DNA sequence elements and influence the preinitiation complex in a way that is not entirely understood (Werner et al, 2005) causing it to be activated or suppressed. Therefore these transcription factors may be either activators or repressors, as is discussed below.

Since the core and proximal promoter together form the promoter, this resultant promoter may be viewed as a region that comprises of two different types of region, one for the polymerase enzyme with the basal apparatus and the second for other elements. These (core and proximal promoters) may be interrupted by spacer sequence or in some cases this division between the core and proximal regions may be less well defined.

In bacteria repressor proteins repress transcription by binding to a sequence that overlaps with the bacterial promoter, thereby inhibiting transcription. In contrast to this, in eukaryotes the chromosome structure itself may inhibit transcription in a general sense. In

eukaryotes it seems that activation of genes is more important than their repression, since the chromosomal structure (at least in part) acts to repress transcription.

For activators or repressors of transcription that interact directly with the PIC, there is a question as to how they are able to contact the core promoter often at long distances from it. Looping of the DNA is thought to permit this. This would involve two or more activators or repressors binding to the DNA and then to one another to form a loop (Rippe et al, 1995, Celniker et al, 2007).

Activators

So how do transcriptional activators operate? The different modes via which the activators may function are as follows: Firstly, they interact with DNA at specific sequences and also with general transcription factors at the core promoter. Many activators also interact with coactivators which do not bind DNA, but assist the activator. They often act in a tissue specific manner. Typically, they have a DNA-binding motif and also an activation domain (Ptashne et al, 1997). The activator protein may for example, aid the RNA polymerase enzyme and its auxiliary proteins to bind the promoter sequence (Wu et al, 2006) and therefore to increase PIC assembly (Li et al, 1999). This constitutes the recruitment model for gene activation

For example, in yeast Gal4 activates genes for galactose metabolism via binding of this activator to its DNA binding site upstream of the core promoter; it is thought that this protein helps to recruit the holoenzyme to the DNA. When other genes are modified so that their upstream region contains the Gal4 DNA-binding site, transcription does take place. Gal4 is also functional in higher eukaryotes.

There is evidence that in Gal4 the DNA-binding region is separate from its activation region. This has been demonstrated by artificial PIC recruitment experiments (Ptashe et al, 1997). Here a subunit of the activator is fused to a DNA-binding domain (which is bound it binding site) and is found to activate transcription.

These types of fusion protein direct PIC assembly and demonstrate the function of the activation domain of the Gal4 transcription activator. Further work (Bhaumik et al, 2004) with Gal4 has shown that the activator domain of Gal4 does not act directly on the PIC, but rather that other proteins or complexes interact with Gal4 which in turn interact to recruit the PIC.

A second mechanism of activator function is in promoting either initiation or elongation. This is presented as a second type of mechanism since it would be a process that happens after PIC assembly. A third is by recruiting chromatin modification. In this case the activator may help to recruit enzymes that can chemically modify the chromatin, for example by directing histone acetylases to the region of a specific gene (Edmondson et al, 1996). Acetylases or methylases can reduce the tightness of packing of the DNA around genes allowing for gene activation in particular locations. Three classes of protein may be associated with the RNA polymerase holoenzyme, that are involved in chromatin remodelling. These are; 1. Histone-modifying enzymes, 2. Chromatin binding proteins, 3. ATP-dependent nucleosome-remodelling proteins (Hamsey et al, 1998).

Despite this though, there is evidence to suggest that holoenzyme recruitment to the core promoter provides the capability to overcome any suppression to transcription that may be posed by nucleosomes. The implication being that chromatin remodelling is more of a general rather than specific form of transcription repression.

For example, in yeast the at the PHO5 promoter histones are modified by transcription activation. This gene is activated by PHO4, an activator with a DNA specific binding site at the PH05 proximal promoter. When this activating region is substituted for a yeast holoenzyme component, chromatin is remodelled and transcription occurs (Gaudreau et al, 1997). This suggests that only holoenzyme recruitment and not a gene specific activator is needed to remodel chromatin. This idea is indirectly strengthened by the observation that histone depletion results in the automatic transcription of many genes.

Repressors

Repressors may be divided into two categories; 1. Global and 2. Gene specific (Gaston et al, 2003). Global repression would result in down-regulation of all genes which are transcribed by PolII, since this type of repressor protein would modify a PIC component. Nucleosomes may also act as global repressors since they form chromatin at the promoter region.

Also, in eukaryotes methylation appears to be connected to transcriptional control, with 2-7% of cytosines are methylated, mostly on CpG/GpC steps. Active genes are relatively under-methylated, and methylation at the promoter can prevent gene expression (Singal et al, 1997). Both Nucleosome modification and DNA methylation present more large-scale repressors of rather than specific regulators of transcription.

Gene specific repression refers to suppression of transcription of a particular gene or set of genes. This is achieved by either decreasing the presence of an activator at the promoter or alternatively by interacting with PIC components or in yet another manner by recruiting chromatin remodelling proteins.

In general, repressors bind to the upstream DNA and help to regulate transcription either close to the proximal promoter or at a distance from it (Gray et al, 1996). This is referred to as 'long-range' or 'short-range' repression. In short-range repression the repressor may block the function of DNA-bound activators in close proximity.

If the repressor acts at a distance, it may contact a PIC activator by looping the DNA in order to make long-range contacts. Also, within the upstream region of a given gene, the long-range repressor will cause the promoter to become resistant to all enhancers including long-range enhancers. Long-range repressors may also act more generally to silence an entire chromosomal locus.

Repression of transcription that takes place via the basal machinery may occur via the following mechanisms. Repressors may either bind to or modify PolII or the general transcription factors. This blocks binding to the core promoter. Repressors may disrupt TFIID binding to the TATA-box either by binding to TFIID or by binding to the TATA-box. Alternatively, they can inhibit general transcription factor interactions or block activators from interacting with the general transcription factors.

Examples of gene specific repressor proteins that bind to the DNA are; Eve (Austin et al, 1995) and MeCP2 (Lewis et al, 1992). Eve is a sequence specific DNA binding protein whereas MeCP2 is a methyl-CpG binding protein. Eve proteins are thought to bind to low-affinity DNA-binding site around the TATA-box, thereby inhibiting TBP binding and disrupting PIC assembly.

Finally in the human positive regulatory elements are usually present at around 50-300 bp upstream of the TSS, whereas negative elements are found at 500-1000 bp upstream of the TSS (Cooper et al, 2006). This is within the approximate proximal promoter region and may highlight a general division or boundary within the proximal promoter.

1.1.4 The enhancer / enhanceosome and the repressor /repressosome

So far activators and repressors of transcription have been described and their likely approximate location within the proximal promoter has been indicated. Now the activity of activators and repressors will be discussed with respect to enhancer elements (and the enhanceosome) or repressor elements (and the repressosome).

The enhanceosome

Eukaryotic enhancers are made up of multiple transcription activator protein DNA-binding sites that operate together in a synergistic manner (Merika et al, 2001). These activators are thought to interact either directly or indirectly via coactivators with the PIC. The enhanceosome may be defined as the combination of enhancer DNA together with activators and coactivators to form a nucleoprotein complex.

Enhancers are often distant from the promoter (Sipos et al, 2005). The activity of the promoter may be greatly increased by an enhancer. The enhancer is made up of a group of elements and is less fixed in space than promoter elements and can often function in either orientation (Tsai et al, 2000). It is unknown if all protein coding genes possess enhancers and also how many enhancers there may be in any given gene. Despite the fact that enhancers can function at varying locations, perhaps within the genome they have a tendency to be arranged in a particular way or clustered according to some unknown system?

An excellent example of enhanceosome is the human interferon-beta gene (IFN-β), which is virus inducible (Panne, 2008). Transcription of IFN-β requires DNA binding and activation of the following transcription factors; ATF-2 / c-Jun, IRF-3, IRF-7 and NFκB (p50). The enhancer element in the promoter of IFN-β is 47-to-102 bp upstream of the TSS. The transcription factors/activators bind to this region to form an enhanceosome. These act synergistically and individual transcription factor binding at this enhancer element is not sufficient for gene activation. The enhancer element contains four regulatory domains that are able to interact with the coactivator CBP. Once this enhanceosome is assembled nucleosome acetylation and chromatin remodelling occur which provides access of TBP to the TATA box. Transcription activation thereby occurs.

The repressosome

In the same way that activators and coactivators together with an enhancer element can form an enhanceosome, repressor proteins together with a repressor element may form a repressosome (Courey et al, 2001). One example of long-range repression is seen with the drosophila Groucho protein. This is a corepressor that does not bind DNA directly forms part of a repressosome. This is thought to occur in the zen gene, which contains an upstream silencer or repressor region. This region contains multiple transcription factor binding sites including, Dorsal, Dead ringer and Capicua (may be an architectural factor). These are thought to act together when bound to the DNA repressor region and recruit Groucho. Groucho in turn blocks PIC formation at the core promoter.

1.1.5 Cross interactions and boundaries in the upstream region

Interactions between the upstream elements: Mediator

The pre-initiation complex interacts with various regulatory proteins. An important example is the Mediator complex (Lewis et al, 2003). This has been found in yeast and mammals (Kornberg, 2007) and is composed of approximately 20 subunits. Mediator interacts with activator proteins and also with PolII, acting as a mediator, hence its name. It is thought to be required for transcription and according to some opinions it is no less essential for transcription than PolII.

Mediator may act as either a positive or negative regulator and its role is to transfer information from enhancers to the core promoter, thereby transferring information from activators or repressors to PolII. Mediator does not support basal transcription but instead it promotes activated transcription. At the same time Mediator may be considered as part of the PIC (pre-initiator complex) and it (or at least some of its subunits) is necessary for non-basal transcription of almost all genes (Conaway et al, 2005). The exact mechanism of Mediator activity has yet to be determined.

Boundaries between core promoter, proximal promoter and enhancer

The overall architecture of the upstream region is highly variable within the human genome and is a subject that requires much more investigation. The following is an outline of some of the issues that pertain to this area. When referring to core promoter, proximal promoter and enhancer (or repressor) elements physical, functional and conceptual boundaries may be considered. It is important to note that there are some grey areas. In other words, in reality the three regions act in unison and boundaries between the core promoter, proximal promoter and enhancer/repressor are not always clear.

It is the combination of interactions of the core promoter with the proximal promoter and the further (or sometimes very distant) upstream elements that result in gene activation and more specifically effective tissue specific transcription at the required level. The general principles may be described as follows;

- The core promoter is required for basal, low level transcription. It contains, locates and initiates the transcription machinery.
- The proximal promoter determines transcription in response to biological signals in a tissue and temporal specific manner.
- The enhancer and repressor are likely to determine the level or rate of transcript produced. It is possible that these elements serve to finely tune levels of transcript via the various binding motifs that it may contain. The functional difference between these elements and the proximal promoter are unknown.

One way to define the promoter (core and proximal together) is as a group of sequences or elements that are ordered in a particular way and are in a relatively fixed location with respect to the TSS. This type of definition excludes any distant enhancer or repressor region since it does not have to be fixed in its location in order to function.

Like promoters, enhancers are modular. There are elements that may be found in both. There have been certain elements found within proximal promoters that resemble those of enhancers in that they are able to function in either orientation and at varying distances from the start site (Huang et al, 2003). This is indicative of the grey area referred to above, regarding the distinction between the proximal promoter and enhancer. In a sense the enhancer (and repressor) can be viewed as being part of the promoter but at a distant location.

The next issue is that of actual physical boundaries and the sequence that occurs between the regulatory elements. Much of this DNA regulatory sequence is thought to be

separated by spacer sequence. The possible function of these spacers is unknown although they may play a structural role and may be involved in packaging the DNA into chromatin.

The interspersed regulatory sequence with spacer sequence makes eukaryotic gene regulatory sequences difficult to identify. This observation begs the following question: What could be the possible benefits of regulatory sequences occurring so far upstream as is seen with distant enhancers or repressors, when presumably they could be placed closer to the promoter? Also, is the boundary and spacer between the promoter (or enhancer/repressor) elements and the intergenic spacer sequence defined by characteristic structures or sequences?

It has already been mentioned that the proximal promoter (at least in some cases) possesses a likely region for positive regulation as well as a potential region for negative regulation. Therefore what is the difference between the positive proximal enhancer and the other more distant upstream enhancer? Indeed in the literature there is reference on occasion to a 'proximal enhancer region'. The difference between the 'proximal enhancer region' and the non-proximal enhancer would appear to be in physical location or distance, since there are enhancers that are present up to many kilobases upstream of the TSS.

What then is the significance of this distance? This issue will only be addressed by further research into this field. It may be for example, that in the eukaryotic upstream region of a given gene there are many (more than are presently identified) enhancer and repressor elements that are available for response to varying physiological/developmental signals. Their arrangement along the upstream region and relative distance from the promoter may be dependent upon their requirement to negate a response to a previous signal.

If basal transcription occurs at a very low level and also other mechanisms exist for gene silencing such as chromatin formation etc..., why are repressors necessary? This remains an unanswered question. However, in theory a repressor may be needed in order to counteract an activator so that a gene is conditionally switched. This would constitute an embedded process.

For example, a set of genes named *SETA* is switched on by an activator in response to a biological signal called *signal1*. Now a second related biological signal (*signal2)* is triggered. This *signal2* causes a subset of *SETA* (called *SETA1*) to be switched off. For this a repressor is required. Simultaneously the remainder of genes within *SETA* (called subset *SETA2*) remain on. This is because the repressor only acts on *SETA1*. This allows a subset of genes to be conditionally switched on or off. In order to validate this type of theoretical scheme or identify alternatives further research is required.

1.1.6 Chromatin structure; influence on transcription

In eukaryotes, DNA becomes greatly compacted (by a factor of about ten thousand) to form chromosomes. It may be that this chromosome compaction controls access of the RNA polymerase enzyme to the start site of genes (Dilworth et al, 2001, Chen et al, 2006). The nucleosome is therefore a general repressor of transcription activity. It has a general 'masking' effect on the DNA.

Conversely, transcriptionally active chromatin possesses an open structure. When transcription occurs there is a change in chromatin structure at the promoter making the DNA temporarily available for the recruitment of proteins that make up the transcription machinery. The nucleosome is found to be absent from active promoters. This has been deduced from reduced densities of core histones at promoters of active genes (Lee et al, 2004) Transcription factors bind to both promoter and enhancer and these may recruit chromatin remodelling enzymes and histone-modifying enzymes which alter nucleosome positioning.

The HMG domain proteins are architectural proteins that are non-histone components of chromatin (Grosschedl et al, 1994). Some of these proteins contain many HMG domains that have low binding specificity for DNA whilst others have a single HMG domain that recognises specific DNA sequences. These are often sequences that possess unusual helical structures. HMG proteins are able to bend DNA and are thought to aid the formation of nucleoprotein complexes. It is within this context that they are thought to play a role in transcription where such complexes are involved in regulation.

A higher level chromatin structure exists which may be involved in gene regulation. DNA is thought to form loops (different to the loops previously described) at regular intervals along the chromosome. These may be approximately fifty thousand base pairs in length. This would mean that each loop contains about fifty turns of 30Å fibre or 250 nucleosomes. The most direct evidence for these loops is seen in 'lampbrush' chromosomes of frogs or newts (Angelier et al, 1990). This is observed in cells that are in the process of becoming egg cells and are therefore producing much DNA. Microscopy studies show RNA polymerase packed along the loops. RNA polymerase is thought to travel around the loop. This looping of DNA may contribute greatly to how genes operate.

It is thought that these loops each contain a gene. If this is indeed true it would seem that this may be the cell's way of separating the genes into discrete units so that they should be recognised as such by the transcription machinery. Since it is known that transcription factors

may regulate gene expression at very great distances from the TSS (by looping for example), what would prevent an enhancer from affecting the preinitiation complex of a neighbouring downstream gene? This large-scale looping would therefore provide a way of demarking individual genes preventing cross-interactions. Therefore all-in-all it is evident that the control of gene expression and in particular transcription is very complex in eukaryotes.

1.1.7 Networks of Gene Expression and Combinatorial Effects

An important question in biology is how the regulation of eukaryotic genes is networked. It is the partnership between the regulatory proteins and their DNA-binding motifs within different sets of promoters that is a key to understanding this issue. In the human genome there are an estimated 30,000 genes (Claverie, 2001); although there could be up to seventy thousand.

If each gene had its own unique activator/repressor, which would specifically bind to the DNA near the TSS, at least half of the genes would be required to code for regulatory proteins alone. Clearly this would seem to be unrealistic since it is inefficient and wasteful. In fact, about 7% of human genes are estimated to code for regulatory proteins (Brivanlou et al, 2002). It may be possible that alternative splicing generates some added diversity for the transcription factor complement, although it is unlikely that this would be sufficient to cover the requirement for an individual transcription factor for each gene. Moreover, in reality there is far more than one transcription factor that binds to the proximal promoter region of any one gene.

Also, if each gene were controlled by a separate transcription factor, it would remove the possibility of groups or hierarchies of transcribed genes. Several (or many) genes are often required to operate together for a given biological process. Examples may include certain housekeeping functions such as the expression of histone proteins or alternatively proteins required for mitosis to occur. Therefore the regulation of genes is likely to be clustered in groups, according to the cell's particular requirement.

In the human (and other higher eukaryotes), transcription regulation relies on multiple biological signals. Different combinations of gene regulatory proteins are expressed in different cell types at specific times and therefore define the unique characteristics of each cell type. In order to finely tune transcription so that it is tissue specific and temporally and developmentally relevant, there is likely to be the potential for multiple activator/repressor to core promoter interaction within any one gene.

Binding sites to these activators/repressors present in the upstream region of the gene would then act synergistically. In some cases the DNA-binding sites and the activator proteins that bind them would be present in the promoter of a large set of genes and in other cases they may be much more specialised. This binding site presence would depend upon the requirement of the activator/repressor in the target 'gene set' which in turn is dependent upon physiological or developmental processes.

It is likely therefore that sets of genes are expressed together. This is despite the fact that they are each expressed from individual promoters and may not be arranged as neighbouring genes. In fact in the human they are likely to be distant and even present on different chromosomes. There is also an added complication in that for a given biological process the levels of required individual protein product may be unequal and utilised in tandem rather than all in one go. Therefore transcription factors may act loosely as members of a group. This would involve the existence of many 'modules' that result in the use of different combinations of regulatory elements and the proteins that bind to them.

Response elements are promoter elements bound by an inducible transcription factor that identify genes that are under common regulation. An example of response element activity can be demonstrated with the human glucocorticoid receptor protein (Eriksson et al, 1995). During times of starvation this hormone is released in the body. It stimulates liver cells to increase glucose production from amino acids. The liver cell must increase the expression of a combination of different genes for this to happen. The combined expression relies on the binding of the glucocorticoid receptor complex to a regulatory site in the DNA of each gene and therefore regulates all of the required genes.

Each of the genes in such a co-regulated set contains its own regulatory region that in theory (although not necessarily) could possibly be involved in other pathways when under the influence of a different regulatory complex. This would mean that these genes or subset of them could be involved in more than one biochemical pathway and would therefore be regulated by different responses.

1.1.8 Transcription factors and their DNA binding sites

A very important issue in biology is the mechanism of transcription factor interaction with and recognition of DNA. This must be specific to an extent since the protein

distinguishes its binding site across the whole chromosome, and acts on a particular stretch of DNA that pertains to the gene in question. This process is not well understood. The question of how transcription factors recognise binding sites and a potential recognition code has important implications for understanding gene regulation and also for drug development.

Protein-DNA binding interactions: docking and probing

Protein-DNA binding involves two phases which are known as docking and probing (Calladine et al, 2004). Docking is the overall fitting together of protein with DNA. Probing is the bonding between the protein and DNA at contact sites on a smaller scale. This detailed probing is successful only if there are specific binding sites on the DNA, which in turn depends on DNA sequence.

Firstly the amino acids fit together precisely in the protein particle in such a way that it must complement the surface of the DNA. Then both surfaces have to match with respect to the different hydrogen bonds and hydrophobic contacts. During bond formations (between protein and DNA) energy is released, however in order for protein-DNA binding to occur an energy barrier must initially be overcome. The overall process must be energetically favourable for the protein-DNA complex to form. Altogether, these protein-to-DNA bonds result in the specificity of the DNA sequence in preference to other multiple sequence stretches on the DNA to which the protein may be able to dock.

It is thought that on average about two-thirds of the contacts between protein and DNA are due to a close fit of their surfaces. Less than one-third is a result of direct hydrogen bonds (Luscombe et al, 2001, Ahmad et al, 2006). The direct recognition of amino acids via these hydrogen bonds (H-bonds) at the base-edges of the DNA is referred to as 'direct readout'. The other type of recognition is called 'indirect readout' which results from the local twisting and curving of the DNA and involves most phosphate-backbone contacts. Indirect readout is the major contributor to protein-DNA binding, although conventionally it is not considered to be the major contributor to specificity of binding. This will be discussed in the text that follows together with implications for a protein-DNA recognition code.

DNA and its structural and chemical features

DNA has a negatively charged sugar-phosphate backbone and base-pairs that are exposed on its major and minor grooves. H-bond donors and acceptors exist on the on both of the grooves. This produces different potential H-bonding patterns for protein-DNA interactions. The result is that within different sequences there are variations in the potential H-bond donors/acceptors and this constitutes a chemical code.

Seeman et al, 1976, attempted to deduce a code for the recognition of DNA by proteins via the availability of hydrogen bond donors and acceptors at the base edges of the DNA. They used these acceptor donor patterns to predict the likelihood of amino acid recognition of the bases. Whilst some of these interactions have turned out to be favourable there has not been identified a code that specifies a one-to-one amino acid base interaction.

This chemical code is superimposed upon overall helical structure that is also determined by the base sequence. It is these together that are ultimately recognised by transcription factors. For a more in depth discussion of the H-bonding donor/acceptor patterns see the introduction of chapter3 (section 3.1.2) and for details on structural properties of the helix see the introduction of chapter2 (section 2.1).

Levels of specificity transcription factor binding to their DNA target sites

Whilst no universal rules (i.e. for all protein-DNA interactions across all transcription factor families) have yet been identified for amino-acid base specificities, certain interactions show relatively strong favourability (Luscombe et al, 2001 and Mandel-Gutfreund, 1995). In general hydrogen bonds show more specificity of interaction between amino acid and base than do van der Waal's interactions.

The following hydrogen bond interactions are most favourable; (i) arginine, lysine, histidine and serine with guanine and (ii) asparagine and glutamine with adenine. Among Van der Waal's bonds; interactions of proline and phenylalanine with thymine and adenine were found to be relatively favourable. However, when considering which amino acid side-chains contact with bases in the protein-DNA complex, it quickly becomes apparent that the side-chains are often ambiguous in their interactions and that there is no coherent code.

Experiments by Luscombe et al, 2002 were carried out in order to investigate the specificity of transcription factor binding to their DNA target sequences. This was done by

investigating the effect of amino acid mutations on binding and also by analysing whether amino acid residues that contact the DNA are more highly conserved than those that do not.

DNA–binding protein families may be grouped into three types according to their level of specificity for DNA sequences, i.e. according to their level of dependence of the base-sequence for binding to happen. These are as follows; (i) Non-specific families where binding does not occur at specific bases on the DNA. Four out of a total of twenty-one families studied fall into this category. (ii) Highly-specific families where contacts with DNA sequences are at similar bases. (iii) Multi-specific binding were contact with DNA is specific but different family members bind different bases.

The vast majority of the protein families (seventeen out of twenty-one) undergo specific binding with the DNA. Results show that in general within protein families, amino acids that contact bases are more highly conserved than those that do not and that within specific DNA–binding protein families there are more base contacts than in the non-specific families. Proteins also contact the DNA backbone and these interactions were found to be conserved in all the DNA–binding protein families.

DNA-binding sites for proteins are usually short and are on average 4-8bp's long, but may be up to 20bp's. Some of these bases within a consensus sequence will vary; often up to half of the bases. This means that effectively a protein may be capable of binding a range of different DNA sequences that are conserved only at some of the base pairs. The conserved bases on the DNA tend to be those that are involved in direct interactions with amino acid residues. This means that protein families that are 'non-specific' and do not form many direct contacts with the bases will tend to be capable of binding many more DNA sequences, hence their so-called 'non-specific' nature.

The contribution of different types of bonds to protein-DNA binding

In order to understand the mechanism of protein-DNA binding it is necessary to know the relative contributions of direct and indirect readout to this binding. Contacts between the protein and DNA that confer specificity are considered to be those that are formed between the amino acid residues and the base edges. Other contacts are less specific. The following is a breakdown of different types of bond formed at varying locations on the protein and DNA together with discussion of these issues.

Of the hydrogen bonds that are formed between amino acids and the DNA many more are formed with the backbone (68%) than with the base edges (32%) (Luscombe et al,

2001). For van der Waal's interactions, it is also the case that many more contacts occur with the backbone (78%) than base edges (22%). A similar relative proportion is observed for water mediated interactions: backbone (71%) and base edges (29%). Of the total protein-DNA contacts made (both with the DNA backbone and the base edges) the following relative proportions were found; H-bonds (20%), van der Waal's bonds (65%) and water-mediated contacts (15%).

This means that van der Waal's bonds constitute the majority (2/3) of protein-DNA interactions and for the most-part they are related to the overall docking of the protein to the DNA. This type of interaction to a large extent is likely to involve overall fit and stabilisation of protein to the DNA and is related to stereochemistry and the geometric properties of both particles. This is related to the process of protein docking to the DNA. Therefore, although there are for example van der Waal's interactions that are considered direct since they exist between the amino acid resides and the base edges they are not necessarily 'direct' in the same sense as H-bonds. This is because the van der Waal's bonds do not form with the base-edges at locations that appear to constitute a potential bonding pattern as exists at the H-bonding donor/acceptor locations of the bases.

In a conventional sense direct readout is considered to be more sequence related since direct contacts are formed between the amino acids and base edges as opposed to indirect readout where contacts are formed between with the backbone. However, if bonds between amino acids and the phosphate backbone are dependent on geometry, then it may be that even these interactions are to an extent sequence specific. This is because the overall geometry of the helix is dependent on the base sequence and also protein (tertiary or quaternary) structure is ultimately dependent upon primary structure. The rules though that govern this 'geometric' sequence specificity may be different to those that govern conventional direct readout specificity.

Mandel-Gutfreund et al, 1995 carried out an analysis of all H-bonding interactions between regulatory proteins and their DNA binding sites. There are four types of possible H-bonds that can form between the protein and DNA and their distribution were as follows; (i) protein backbone and DNA-backbone (18%), (ii) protein backbone and DNA-base edge (1%), (iii) amino-acid side chain and DNA-backbone (51%), (iv) amino-acid side chain and DNA base edge (30%). Therefore of all possible H-bonding interactions between the protein and DNA, a minority (only 30%) can be considered truly 'direct', .i.e. those between amino-acid side chains and DNA base edges. These are the interactions that are thought to play a key role in specific sequence recognition.

Indirect interactions occur through the backbone and although in total the proportion of H-bonds formed are greater with the DNA backbone; these interactions are generally considered more indirect and are thought to have less influence on specific protein-DNA sequence recognition. Instead they are thought to stabilise the protein-DNA complex. Also, it is for direct binding interactions, i.e. H-bonds between amino-acid side chains and the DNA base edges that the most significant interaction relationships were found. For example, guanine often is involved in such interactions and its H-bonding with arginine and lysine is statistically higher than the expectation (Mandel-Gutfreund et al, 1995). This type of relationship was also observed for H-bonding interactions with other amino-acid resides and base-edges. In contrast, indirect interactions appear less significant, i.e. more randomised.

Pabo et al, 2000 have studied geometric characteristics of protein-DNA complexes. This analysis was carried out in a manner that is independent of the actual identify of the base or amino acid. Instead spatial relationships were considered. This is advantageous since it allows position and orientation of the two particles with respect to each other to be considered which is useful for determining which contacts are likely to be made. In contrast, in an analysis of contacts alone these spatial or geometric relationships are omitted.

The results of this study showed that when similar protein motifs have similar amino acid to base interactions, the spatial or geometric relationships are conserved. When the same amino acid contacts different bases (for example within different locations in a complex) the geometric relationships are usually very different. This means that spatial factors can dramatically alter the bonds formed between amino acids and bases and may explain the ambiguity regarding observed amino acid-base contacts. Although within a given protein family there may be some characteristic contacts (see above) there is actually is a relatively high variation in the geometric arrangements within the same family.

This type analysis of protein-DNA interaction is one that pertains to protein docking to the DNA since it is this step in the interaction that relies on fit or geometry. It serves to highlight the importance of docking in protein-DNA recognition. This would be important for all protein families that bind DNA. However, for those protein families that are considered 'non-specific' in their binding to DNA (see above) where binding does not occur at specific bases on the DNA and therefore they are capable of binding multiple DNA sequences, this factor of geometry may be a greater determinant of binding than protein families that are specific.

Both docking and probing–related interactions are likely to be sequence specific. Even docking is probably determined by some level of sequence specificity since it is related to geometry which is ultimately reflected by sequence. However, the sequence determinants for each are likely to be very different and are certainly not straight-forward. This means that the

relationship between sequence and docking interactions is not necessarily the same as the relationship between sequence and probing interactions.

Different classes of transcription factor

Transcription factors contain DNA-binding sites which often possess globular domains that interact with DNA. This is the location on the protein that 'reads' the DNA and therefore its nature is of great importance for protein-DNA binding. There are though other locations of the protein which are also important for interactions with the DNA since they contact the phosphate backbone.

Transcription factors fall into different families depending on the way in which they interact with the DNA. This means that the 'reading' motif of proteins (as well as other contact sites) can be divided into categories or families. Therefore with regard to the recognition of DNA by the protein there are likely to be some general rules that pertain to all transcription factor binding and also some family-specific rules for binding.

There are different ways in which to classify protein folds that bind to DNA. Luscombe et al, 2001 subdivided transcription factors into thirty different classes and this subject of different types of protein fold have been reviewed by Garvie et al, 2001. It is usually α-helices or β-sheets of transcription factors that make contact with the DNA by penetrating the DNA grooves. The α-helix usually contacts the major groove although it may also make contact with the minor groove. Several examples of DNA binding folds will be briefly outlined and then the zinc finger transcription factor, Zif268 and its DNA binding site will be discussed in more detail.

The helix-turn-helix motif contains two α-helices with a bend between them, with the helix length varying amongst different members of this general class. One of the helices contacts the DNA by inserting into the major groove and forms contacts with the base edges. An example of the helix-turn-helix motif is found in the lambda-repressor. The basic region-leucine zipper and helix-loop-helix proteins are dimeric proteins. These two classes of transcription factor bind DNA in similar ways. Here the N-terminal region of the α-helices insert into the DNA. Examples include; E47 and MyoD.

Transcription factor α-helices can in some cases also contact the minor groove of the DNA. However, the DNA is bent or distorted, as for example is the case with PurR, the purine repressor dimer. The zinc finger family of transcription factors is the most prominent in the human. Here the DNA binding domain is comprised of a short α-helix, two anti-parallel β-

sheets and a core Zn^{2+} ion. Zinc fingers contact the DNA by inserting the α-helix into the major groove. β-sheets are able to form DNA-binding sites in some transcription factors, although their occurrence is relatively rare. Some bind the major groove whilst others bind the minor groove. TBP binds the minor groove of the DNA by inserting a 10-stranded β-sheet and causes a profound distortion of the DNA helix.

In eukaryotes there is are a far greater variety of transcription factor DNA binding motifs than in prokaryotes and therefore the eukaryote is enabled in more diverse ways of interacting with the DNA. The total complement of transcription factors constitute the main executers of the program that develops and maintains the organism, therefore the more complex the organism the greater the diversity of its transcription factors. Also, in eukaryotes the proteins that bind the DNA are often asymmetric as is their interaction with the DNA. In contrast in the prokaryotes transcription factors tend to be symmetric and dimerised. The reason for this difference between the prokaryotes and eukaryotes is unknown. However, the asymmetric transcription factors may constitute a higher level of variety of protein-DNA interaction.

Zif268 and its DNA binding site

So far the best information about specificity of transcription factor to DNA binding lies in X-ray crystal data, which has yielded structures of different protein-DNA complexes at near atomic resolution. One example; Zif268 will be given in order to illustrate the complexity of protein-DNA binding (Jamieson et al, 1994). This is a member of the zinc finger family of regulatory proteins (described above).

Zif268 binds to the DNA via three successive zinc fingers that are connected to each other by a flexible peptide linker. Each of these zinc fingers inserts an α-helix into the major groove of the DNA helix. These zinc fingers bind to a regulatory element within the DNA sequence and interact with specific bases (see figure 1.2). The amino acids make contact with bases in both strands of the DNA. In some cases only one base in the pair is contacted by the protein, whilst in other cases both bases in the pair have contacts with the protein. However, in the case of Zif268 the binding sites of neighbouring fingers overlap, with each of the bases in a base-pair interacting with a different zinc finger.

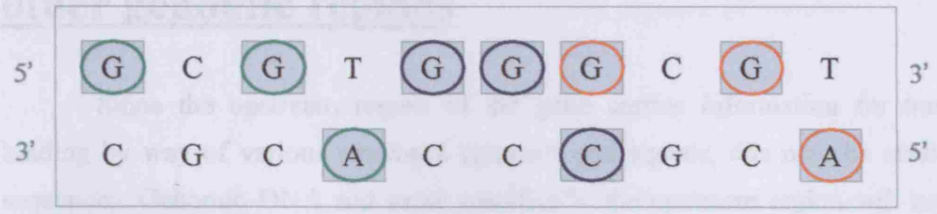


Figure 1.2 Diagram of transcription factor binding site for Zif268
This is the regulatory motif to which Zif268 binds. The protein contains three zinc fingers. The amino acids side chains of these zinc fingers bind to the bases within the above-given sequence. Bases shaded in grey are those with which there is direct binding with the protein. The bases encircled in green, blue and orange each bind to a different one of the three zinc fingers. This binding pattern is complex, with different bases in the base-pair being bound by different zinc fingers of the protein.

From this example it is clear that protein-DNA recognition and binding is complex. Interactions between amino acid side chains and the bases or phosphate backbone may occur. Also, the pattern of interaction along the sequence of the DNA binding motif cannot be predicted according to sequence. For example, interactions may be with either of the bases in the base-pair, some bases may not be involved in direct bonding interactions, and interactions may be via hydrogen bonding or hydrophobic interactions.

Is there a recognition code?

The much sought-after 'recognition code' via which protein recognition of DNA occurs remains non-deciphered. Therefore how exactly a protein recognises and reads the DNA sequence and selects its target from many other potential sites is unknown. It has been found that in general the interactions of amino acids with base pairs are not of a 'one-to-one kind' nature. They seem to be context dependent in such a way that is very difficult to predict (Benos et al, 2002, Pabo et al, 2000). This in turn makes it difficult to predict the DNA regulatory elements, their arrangement and their potential to form networks of gene regulation of the type previously discussed. Zinc finger proteins however, have been found to contain some level of recognition code (Suzuki et al, 1994, Elrod-Erikson, 1998), although this is very limited.

1.2 Sequence characteristics of the upstream relative to other genomic regions

Since the upstream region of the gene carries information for transcription factor binding by way of various structural elements and signals, this may be studied via the DNA sequence. Genomic DNA and more specifically the upstream region will now be introduced with respect to sequence features and a general outline will be presented as to the importance of DNA sequence and what may be learnt from it.

1.2.1 General sequence characteristics of genomic DNA

Genomic DNA has non-random sequence characteristics (Blaisdell et al, 1883, Karlin et al, 1993) and also possesses a pervasive and stable signature that varies from species to species (Karlin et al, 1997). These characteristics and the non-randomness of DNA sequences are discussed later on in this chapter and are also dealt with in depth in the introduction to chapter2. In summary though, general sequence characteristics of DNA may be described in terms of a tendency for certain periodicities and for the formation of particular motifs. These in turn are related to structural features; in that DNA has certain helical properties which are defined by sequence. Also, another level is the language-like property of the DNA that is generated by codes such as the triplet code because embedded within these is biochemical meaning.

DNA sequences in general possess a periodicity that repeats itself every ten bases which is the approximate length per helical turn (Li et al, 2006). Superimposed upon this are additional features depending on the sequence type. For example, certain non-coding regions may possess short sequence stretches that are responsible for maintaining the coiled-coil structure of chromatin and nucleosome assembly. Also, the nucleosome core particle consists of a 146-base pair strand of DNA in association with a histone octamer.

This is an example of a tendency within the DNA to form certain motifs. This DNA motif is a location to which particular types of protein (namely histones) bind. These motifs are likely to have characteristics that are either chemically and/or structurally compatible for histone interaction. Local DNA sequence affects helical structure which determines function, an issue that is dealt with in depth in chapter 2, section 2.1.

Superimposed upon the 'background' genomic DNA sequence characteristics (such as helical periodicities) are additional sequence properties that are likely to define particular

regions and distinguish between them. The following text provides a description of both these background characteristics and region specific characteristics.

1.2.2 The genome and its different structural and functional locations

The human genome may be divided into different parts or regions. These possess different functions, despite their close spatial proximity to one another and their sharing of borders or boundaries. Examples include coding and non-coding regions. A further subdivision within the non-coding category is intergenic and intronic regions. Therefore DNA may be said to have not only an inherent general structure, but also varying local structure depending on the region in question.

Since it is a general principle that function is dependent on structure which in turn is reflected in sequence, these different genomic locations are likely to possess characteristic sequence properties. This means that there is a tendency to form different 'words' or motifs at varying functional locations. With regards to coding regions, these 'words' are the 64 different possible triplets, 61 coding for amino acids plus 3 stop codons. Since within the non-coding regions the actual 'words' are an unknown entity they may be regarded as a free parameter. Another way to view this is that there is tendency for the formation of different combinations of base sequence to occur in different regions of a chromosome.

1.2.3 Coding and non-coding sequences

The coding sequence possesses meaning that is ultimately contained within its protein product. In the coding sequence there is a bias for certain motifs since the sequence codes for protein and thereby reflects peptide motifs (Gao et al, 2005). Also, there are codon biases within these sequences (Karlin et al, 1994). This bias refers to the presence of certain nucleotides at particular codon positions and also nearest-neighbour tendencies. The authors hypothesize that one major influence on codon usage in the human is residue preferences in proteins and amino acid constraints.

The upstream non-coding sequence contains regulatory elements (core promoter, proximal promoter, enhancer etc...) interspersed with so-called spacer sequence of varying length. The regulatory regions must be 'read' and recognised by proteins and therefore contain a

distinct 'language' or 'code' that enables this to take place. This so-called language is also likely reflected within the DNA sequence.

It is also possible to think of the upstream sequence in terms of higher and lower functionality regions. Regions that apparently have no function would be of a lower level of functionality, such as the spacer sequence. These may be regarded as being less developed than higher level sequence such as the regulatory sequence. Therefore the spacer sequence may be regarded as the generic sequence which is less meaningful. This is a subject discussed in more detail in the introduction to chapter 4.

Genomic DNA contains repeats. In fact more than 50% of human genomic sequence constitutes repeats. In general the longer the life cycle of the organism the greater the repetitive content of its DNA. This has for many years been considered to be 'junk' DNA since it was thought to lack a function. However, alternative ideas have arisen regarding the possible functionality of repetitive DNA (Shapiro et al, 2005). From this view-point repeats are essential to the genome and play a role in formatting coding information and also in transmitting this during cell division. For the most part repetitive elements are composed of smaller sub-components or motifs. This indicates some level of structural organisation.

The following are some examples of different repeat classes. Repeats are thought to influence nucleosome positioning, for example VNTR elements, which are very flexible and are thought to have an affinity for nucleosomes. MARs or scaffold associated regions are another type of repeat. DNA transposons such as such as LTR-retroposons are thought to be associated with heterochromatin. Repeats involved in transcription and that may be located within promoters or enhancers or silencers include LINEs and SINEs. A LINE element in an enhancer for instance may act as a transcription factor binding site. LINEs may also play a role in retarding transcript elongations.

1.2.4 Extracting information content from sequence to understand structure and function

Long-range correlations have been found in DNA sequences, Peng et al, 1995. The non-random characteristics of genomic DNA may be demonstrated by random walk or fractal models which can show these long-range correlations. A DNA walk can be represented graphically. The walk is incremented either up ($u(i) = +1$) or down ($u(i) = -1$) for each step, i of the walk. For instance, in the DNA sequence at position i, if there is a purine the walk is incremented up or alternately if there is a pyrimidine is incremented down. In a correlated walk

the direction each step is depended in the previous step implying 'memory', whereas in an uncorrelated walk it is independent.

Non-coding sequences have in fact been found to possess some statistical features that are shared with natural language. This is in the sense that long range correlations have been found in written language, since a random walk model can be applied to these. Long range-correlations have also been found in non-coding DNA. This reveals the potential presence of a 'language' in non-coding sequences. We see from this that it is possible to know in a general sense that sequence possesses non-random characteristics and that it contains meaning. Deciphering this meaning is also possible via the use of sequence analysis techniques.

An important question regarding DNA sequences relates to why there is a four 'letter' alphabet and also why those letters possess their particular characteristics? This is an intriguing question since the letters posses both physical and chemical properties that probably relate directly to their biochemical role.

If we consider the properties of the nucleotides, they can generally be divided into two types; 1. purines/pyrimidines 2. weak/strong. Purines and pyrimidines relate to geometry and structure. Sequence variation of these produced a variety of patterns that relate mostly to DNA structure. Weak and strong base pairs have different numbers of hydrogen bonds between the bases that form the helix. However, more importantly, they contain different H-bond donor/acceptors (see chapter3, section 3.1.2) and therefore sequence variation of weak and strong bases produces different patterns for this chemical code. The 'information' content within nucleotides is encoded in both these H-bond donor/acceptor patterns and in purine/pyrimidine motifs.

Donaill, 2002 has described DNA sequences in terms of its information content and connected this with a parity code. In error-coding theory, a code for which all code-words have the same parity is considered to possess error-resistant properties. This type of code is called a parity code. H-bond donor/acceptor patterns can be expressed in binary notation and this is also true for purines/pyrimidines. Together these two-dimensions of information content give the nucleotide a 4-bit numerical representation. Code-word (nucleotide in this case) parity is either odd or even depending on whether the total number of 1's in its binary representation is odd or alternately even. If all the code-words or nucleotides have even parity, for example, the code itself (i.e. the nucleotide alphabet) is considered to be a parity code.

This is in fact the case for the nucleotide alphabet, which does have a parity code structure. The advantage of this, in theory, is in the number of features that have to be altered in

order to change one code-word to another. For a parity code more than one feature must be changed, making the nucleotide alphabet a robust code, from the coding-theory perspective.

Since the structure of the DNA helix and the chemical code (H-bond donors/acceptors) are recognised by transcription factors, if one wishes to study transcription factor-DNA binding it is relevant to study this process it in terms of these two types of nucleotide characteristics. Both of these DNA are determined by sequence and therefore it is useful to analyse the nucleotide sequence in terms of this information content. This informatic perspective regarding the sequence can then be related back to the structural and functional properties of a genomic location such as the upstream region.

1.3 A Brief Outline of Project Aims

In the human genome, the mechanisms of transcription regulation are complex. There are many unanswered questions about the upstream sequence, its organisation and function. The following are some examples of important and unresolved general biological issues regarding the upstream sequence of human genes:

1. Regulatory sequences may be found far upstream of the TSS in the human genome. The reason for this arrangement and the role of the spacer which can be made up of long sequence stretches is unknown.
2. It is unknown how far upstream the regulatory sequence spans. Included in this is the distinction between promoter and enhancer. The upstream boundaries for regulatory sequences beyond the promoter are unknown. The specific sequence and structural distinctions between regulatory and non-regulatory sequence remain undetermined.
3. Regulatory elements consist of short sequence motifs that may occur by chance in the genome sequence and are therefore difficult to identify and characterise. The mechanism via which inappropriate binding is avoided remains non-elucidated.

In order to begin to address these issues, different types sequence analysis methodologies may be employed. The aim of this project was to gain a better understanding of the sequence properties of the 5' upstream region of the human gene. Bioinformatic techniques were used in order to accomplish this. This project begins with a simple analysis of sequence and builds stepwise to a progressively more detailed and in-depth study of the upstream sequence. Two general strategies were utilised:

i) An analysis of changes in sequence properties at different locations across the 5' upstream region.

ii) A comparison of the 5' upstream to other genomic regions.

Information about sequence properties provides a better understanding of how the cell utilises the upstream region to regulate gene expression. Comparing the 5' upstream with other genomic sequences allows for the deduction of features unique to the upstream. Common features between upstream, intronic and coding regions likely reflect features that are inherent to human DNA.

The differences between upstream sequence close to the TSS and sequence further upstream, i.e. the intergenic region were studied. Sequence trends across the 5' upstream region at different positional locations are likely to reflect changes in the organisation of the upstream and in local function. The 10Kb upstream sequence of each human gene was taken and divided into positional segments (ten equal 1Kb portions) or sliding windows. A comparative analysis of these individual sequence portions was then carried out in order to observe any changes across the upstream. Various analyses were performed which will be described in the sections that follow.

The aim was to build an overall picture of the structure and function of the 5' upstream region of the human gene in order to understand biology from the sequence. A global analysis was carried out across human genes in order to build an 'average' picture of the upstream region. Specific upstream elements or sequences were not of interest. Instead a typical or prototype situation was derived from this large-scale study. This analysis was consolidated with existing knowledge in order to extend the understanding of this region of the genome.

All analyses in this entire project were carried out on human sequences from the NCBI database. Once the upstream sequence was collected these datasets were reused throughout the project. Although the obtaining of these sequences has only been described in chapter2 the identical sequence datasets were subsequently reused later chapters; 3, 4 and 5. It is essential to read the chapter 2 methods section in order to understand the methods of the other chapters. Any programs written for dataset compilation and analysis were done so in the C programming language within the Linux environment.

There are four general parts or experimental categories to this project, each one containing its relevant subdivisions. A general outline of these categories may be described as follows:

1. An analysis of dinucleotide sequence composition and representation:

Changes and differences in dinucleotide composition and representation were analysed. This had implications for structure, directionality and strand asymmetry. Dinucleotide content reflects structural properties of the DNA. Changes in dinucleotide compositional tendencies across the 10Kb upstream sequence of the gene therefore relate to structural differences. These structural properties and their alternations towards the TSS are likely associated with the role of this region in gene regulation. This is described in depth in the introduction and aims of chapter 2, section 2.1.

2. Distance from randomness by comparing real sequences to a random model:

This was carried out in order to obtain a general picture of any changes in structure and relative level of functionality across the upstream. If one sequence type is non-random relative to another, this implies differences in function. Changes in the overall non-randomness of the upstream sequence were analysed across the 10Kb upstream sequence in order to see if there would be changes in functionality towards the TSS.

Furthermore the aim of this experiment was to analyse the relative non-random characteristics of purines/pyrimidines and weak/strong sequence across the upstream region. The purines/pyrimidines property of the bases relates to structure and geometry of the DNA helix. In contrast the weak/strong property contains hydrogen bond donor/acceptor patterns and therefore carries a potential chemical code.

By separating out these characteristics, it becomes possible to study the sequence in terms of its information content. Therefore this was done in order to see whether one type of property was more 'important' than the other in the upstream sequence and if this would change towards the TSS.

The 10Kb upstream was therefore 'translated' or interpreted from the original ATCG sequence into; 1. Purines and pyrimidines (R/Y) sequence and into 2. Weak and strong (W/S) sequence. The term translation is not intended here in its biological context but rather as a conversion for instance of A/G into 'R' or C/T into 'Y' within the DNA sequence and has been used in this project to describe this type of sequence conversion.

The identical distance from randomness profile was applied for these two types of translation of the upstream sequence. It then became possible to see if there was a difference between the non-random profiles of these two translations. A more distant from randomness profile suggested a relative higher level of importance for that particular (translated) sequence property. This subject is described in detail in the introduction and aims of chapter 3, section 3.1.

3. **Sequence similarity using patterns analysis:**

The different positional segments of the upstream (from the 5'-to-3' end) were tested for sequence similarity utilising common patterns. This was done in order to see if there is a change in sequence similarity across the 10 Kb upstream region. The general purpose of this experiment may be described in the form of a question: Across the (different positional locations of the) 10 Kb upstream sequences, is there a change in sequence similarity between 'all' human genes? Therefore this is a study of relative divergence/convergence of sequence. The aim was therefore to extend the study of sequence, structure and function of the upstream. This experiment was carried out as a follow-up of the above distance from randomness profile of the upstream sequence. See chapter 4, section 4.1 for more details.

4. **An analysis of the distribution and representation of regulatory elements (mechanism of avoidance of inappropriate binding):**

The distribution and representation of known regulatory transcription factor binding site (TFBS) motifs was analysed across the upstream sequence. This was done in order to gain a better understanding of the arrangement of regulatory sequences across the upstream. Of particular interest was the question of the mechanism via which inappropriate protein binding may be prevented. This by deduction would provide clues for the interaction and recognition of the TFBS by regulatory proteins.

The essence of this experiment was not to analyse TFBS motif occurrence in the upstream only in terms of the ATCG sequence but rather in terms of the R/Y and W/S translated sequences. This is because in the above-outlined experiments differences were discovered for these (R/Y and W/S) properties that have important implications for protein-DNA binding with respect to docking and probing. Therefore regarding the issue of avoidance of random binding of regulatory proteins to the DNA, this R/Y and W/S analysis for the representation of TFBS in the upstream sequence was essential. See chapter 5, section 5.1 for more details.

Each category of experiment was designed to follow a previous category. The later experiments were the consequence of interesting questions that were raised as a result of the initial analysis. The connection between the different types of experiment will therefore become apparent upon further reading and more detail will be given in each of the experimental chapters.

2. A Dinucleotide Analysis of the 5' Upstream Region: Implications for Structure, Directionality and Strand Asymmetry

2.1 Introduction

The mechanisms of transcription regulation are complex. There are many unanswered questions about the human upstream sequence, its organization and function. The reason for the promoter, enhancer and spacer arrangement is unknown. Also, it is unknown how far regulatory sequence in the upstream spans and why. Included in this is the problem of distinction between promoter and enhancer. The upstream boundaries between regulatory element and spacer are also undefined.

2.1.1 Upstream sequence properties

In the genome there are sequences in very close proximity that possess very different roles. A change in the sequence composition reflects a change in structure and function. It is known that the upstream sequence containing the promoter is functionally different in that it is dense with regulatory sequences. The sequence further upstream may possess some regulatory sequence and also intergenic spacer. The way in which sequence, structure and function are connected across the upstream is relatively unknown.

Any given DNA sequence may be described for Bioinformatic purposes as a string that is made up of four letters that are put together in a particular order. Therefore each letter has a particular position within the string. This sequence string (in general terms) likely reflects an overall structure that is characteristic of DNA.

Nucleotide sequence reflects DNA structure, function and organization within different parts of the genome. The structure (and therefore function) of proteins is coded for by the DNA sequence. This is inherent in the triplet code. There are similar motifs that are repeated in different coding regions, which result in similar local protein structures (Mrazek et al, 1992). These in turn affect function. This tendency for certain motif-types gives the coding sequence its characteristic properties. There are differences in base composition that occur across DNA

sequences and have been shown to be related to factors such as codon composition (Karlin et al, 1996) and direction of transcription (Larhammar et al, 1993).

Non-coding sequences should also possess characteristics (different to the coding sequences) that correlate with their purpose. Regulatory elements may make up modules and various combinations of element may be common to different groups of genes. These protein binding motifs may confer the sequence properties that are characteristic of the upstream. This includes sequence composition as well as other possible sequence properties. For example, it is known that within the promoter the CG content increases and AT decreases towards the TSS (Louie et al, 2003). This may be due to regulatory elements.

2.1.2 Why look at dinucleotide motifs?

In looking at nucleotide composition alone their order (or sequence) is not considered. To do this, motifs may be utilized. The simplest motif is the dinucleotide. There are sixteen different possible dinucleotide motifs. Therefore the set of dinucleotides and their relative composition would be the simplest way of describing any DNA sequence.

DNA (either coding or non-coding sequence) is expected and indeed has been found to possess different properties to a random sequence (Blaisdell, 1983, Karlin et al, 1993). A set of random sequences contain no specific or common structure or function. Any similarity in sequence that they appear to possess would be due to chance. In real DNA sequences certain motifs may be over-represented whilst others are under-represented. This non-randomness of DNA sequences has been seen in long runs of nucleotides and also in dinucleotides (or doublets).

The set of dinucleotide representation profiles may be described as a genomic signature (Karlin et al, 1995, Campbell et al, 1999). This shows how different the real sequence is to a randomized equivalent sequence. The dinucleotide profiles may be used to compare different sequence types, each one possessing its own unique signature. Compositional heterogeneity between and within genomes is a recognized phenomenon (Burge et al, 1992, Karlin et al, 1997). The dinucleotides possess non-random characteristics and their signature is in fact remarkably stable within the majority of genomes. Heterogeneic features include: tandem repeats and universal under-representation of the dinucleotides TpA and CpG in many vertebrates. Low CpG frequency, for example, can be explained by methylation, deamination and mutation, which causes CpG to be converted to TpG or CpA (Sved et al, 1990).

A change in dinucleotide representation across the upstream sequence would reveal a change in the importance or use of that dinucleotide. This suggests differences in structure and function. If the dinucleotide is over-represented in the sequence it is enhanced beyond the random expectation whilst considering the nucleotide composition. If the dinucleotide is under-represented, it is specifically suppressed. Therefore an 'effort' has been made, so-to-speak to change the randomly expected dinucleotide content. This has implications which will be presented.

Trinucleotides and tetra-nucleotides and indeed all short motifs are also representative of sequence, however the dinucleotide is the simplest motif and one for which structural features in the DNA are very well characterised. The dinucleotide also provides a very powerful yet simple sequence analysis tool and it is extremely stable providing the genomic signature described above. It is for this reason that the dinucleotide has been chosen for this study. Mononucleotides may be utilised to describe composition but do not provide a description of sequence and structure. Mononucleotide content for the sequence dataset utilised in this project is shown in the appendix A.1.

2.1.3 Structural implications

The location of any object in space may be defined by six measurements called Euler numbers. These include three co-ordinates and three angles. The location of dinucleotide steps in the double helix however may be sufficiently described by only three Euler numbers due to constraints placed on them by the sugar phosphate backbone. These are the roll, twist and slide angles. Each dinucleotide step adopts a different set of these angles due to different shape and charge of their bases (El Hassan et al, 1995, Packer et al, 2000). The overall effects are varying types of helix. Therefore base sequence influences the helical structure that is formed.

Since dinucleotides provide a basic description of sequence, they can also provide information about structure. The DNA helix may in theory exist in different forms, namely the A, B, and Z-forms. The sequence of nucleotides can define the conformation of the helix and these different potential forms of DNA.

B-DNA is thought to be the typical form of DNA. This type of structure minimises repulsive forces between the charged phosphate groups. B-DNA tends to have 10 base pairs per helical turn. In contrast to this, the A-form of DNA contains 11 bases per turn, and the helix is

wider and flatter than the B-form. This is true for ideal A and B-forms. There are though also intermediate structures as has been observed by X-ray crystal diffraction (Ng et al, 2000, Banavali et al, 2005). A third alternative conformation is Z-form DNA. This produces a left-handed helix with 12 base-pairs per turn.

There is known to be a link between DNA that possesses a high frequency of certain dinucleotides (or runs of that dinucleotide) and the tendency to adopt certain helical conformations (Hunter, 1993). For instance, the Z-from of DNA tends to be adopted by sequences that possess alternating purine and pyrimidine bases (Wang et al, 1985).

These different helical conformations are in turn associated with certain biological functions. For example, it is thought that particular segments of the DNA sequence may be converted from the B-form to the Z-form of the helix and that these have a role in regulating gene expression (Reich et al, 1993).

Structural features have been studied in and around TATA box regions and Inr regions (Fukue et al, 2004, Fukue et al, 2005) in the human and mouse. Such studies are important for understanding the mechanical aspects of protein-DNA recognition relating to transcription regulation. Factors like flexibility, rigidity and curvature may help to explain these protein-DNA interactions. Average flexibility profiles were shown using trinucleotide steps. Here it was found that within the upstream half of the TATA box or Inr, the sequence is more rigid than the downstream portion.

Research by Rozenberg et al, 1998 on the viral E2 regulatory protein and recognition of its target DNA has shown that dinucleotides may actually make up a structural recognition code. This emphasizes the importance of dinucleotides as the most basic structural entity in protein-DNA recognition. This recognition code would then be specified by hydrogen bonds and other interactions. Therefore analyzing changes in dinucleotides across the upstream sequence is useful for understanding general variations in structure.

It is known that in genomes in general there is a tendency for YpY and RpR dinucleotides (these contain either two purines or two pyrimidines) to be enhanced whereas there is a minimization of YpR and RpY (dinucleotides containing one purine and one pyrimidine) (Amano et al, 1997). This arrangement probably minimizes deviation from B-form DNA and maintains structural stability.

Pyrimidine-purine steps are relatively flexible. This means that they are able to adopt two conformations and also intermediary conformations (el Hassan et al, 1995), i.e.

conformations between the A-form and the B-form of the DNA. Also pyrimidine-purine sequences have reduced stability due to reduced base-to-base overlap. Therefore the energy needed to unwind the helix (in which these steps are present at a high proportion) is lower than in other sequences. YpY and RpR steps are in contrast rigid. For example, AA/TT is a rigid step, which can only adopt a narrow range conformation (El Hassan et al, 1995, El Hassan et al, 1996) tending to adopt the B-form of DNA.

Another important structural issue is that most DNA sequences only possess a small intrinsic curvature when in solution. However, when in contact with a protein they can often adopt highly curved shapes in order to bend around a protein. This capability is indeed due to helix flexibility. The flexibility in turn is related to the flexible dinucleotide steps.

Furthermore, it has been demonstrated that nucleosomes and regulatory proteins induce different types of bending of the DNA molecule (El Hassan et al, 1998). The bending induced by regulatory proteins appears to be much more pronounced. This ability to bend may therefore represent a recognition system which distinguishes regulatory proteins and nucleosomes. This difference may be reflected in different types of bending properties of sequence that is responsible for transcription regulation.

Dinucleotides that are composed of two strong bases; CpG, GpC, CpC and GpG (or SpS), fall into a unique category regarding the roll and twist angles and resultant DNA structures. A 'strong' base refers to the potential capability to form 3 H-bonds in a base pair. SpS is a bistable step so it is able to adopt two extreme conformations (El Hassan et al, 1995). These dinucleotides are able to adopt either high slide or low slide conformations but without intermediates. If the sequence consists entirely of SpS steps, this can result in either of two extreme structure the A-form or B-form of DNA but not intermediate structures.

The focus here is the observation of general and more large-scale changes across the upstream sequence. Therefore a sliding-window approach for sequence feature changes was utilized. Dinucleotides were considered since this is the simplest motif for which structural properties have been studied. Also, the dinucleotide is thought to be important for considerations of protein-DNA interaction.

2.1.4 Sequence orientation and strand asymmetry

An additional feature of the DNA sequence is that it possesses directionality or a specific orientation. DNA is read by the transcription machinery in a particular (3' to 5') orientation and the mRNA is synthesised from the 5' to 3' end. This confers directionality upon the sequence. Coding sequence for example, is expected to be very highly direction specific since it contains protein structure information. Parts of the 5' upstream (and intergenic) sequence of genes may also possess directionality since they contain protein-binding elements.

This directionality may also be regarded as another type of non-random characteristic of the DNA. This is because a fully randomized sequence contains equal proportions of dinucleotide directional pairs; XpY and YpX. In a region containing a high density of protein-binding motifs, the directionality is expected to be high. This property of directionality also implies a possible mutation bias whereby for example, the formation XpY is favoured over YpX.

Another different (albeit related) feature of the sequence is strand asymmetry. This has been found in many bacterial genomes (Mrazek et al, 1998) and also in eukaryotes (Niu et al, 2003). Different mechanisms have been proposed for this asymmetry. The first is associated with replication and repair on the leading and lagging strands. The second relates to transcription and transcription coupled repair during which there may be mutation biases and deamination events on the coding sequence strand. A third may be due to codon usage.

Work by Louie et al, 2003, has shown that in the human promoter region, just proximal to the TSS, there is a skew of T versus A and C versus G, which is due to biases between the sense and anti-sense strands. This suggests strand asymmetry at this location. Both strand asymmetry and directionality may change across the 5' upstream sequence depending on the relative location.

2.1.5 The effect of repeats

Repeat sequences could potentially alter dinucleotide composition and representation. This is because repeats have a tendency to form certain motifs that are present many times. Therefore they may feature certain dinucleotides more than others. An added complication is that the density of repeats could vary in different genomic regions (Rajendrakumar et al, 2007).

By filtering out the repeats it is possible to see their effect on the upstream sequence. Often repeat fractions are filtered out by default in sequence analysis experiments such as these. However, repeats are an integral part of the DNA. Within the upstream they may play a structural role and may even be involved in the process of regulation (Iglesias et al, 2004, Iyer et al, 1995). The problem with removing them is that a potentially important component of the upstream sequence is removed. For instance, a significant over-representation of transcription factor binding sites has been found in repeat sequences in the human genome (Stepanova et al, 2005) as well as in the repeat free fraction. Therefore the effects of repeats on the dinucleotide properties of the sequence are shown.

Repeats constitute a large proportion of the DNA (around 50%). They have different characteristics to non-repeat regions that are likely to affect sequence compositions and possibly dinucleotides. It is standard practice in sequence analysis experiments to filter out repeats. Other filters were not applied (for example MARS/SARS) since these sequences were not of specific interest in this project and it is not standard practice to filter these out.

2.1.6 Aims and experimental design

The aim was to build a picture of the sequence and obtain insights into the structure and function of the 10 Kb 5' upstream region of human genes via a study of dinucleotide composition and representation. These would reveal information about structural tendencies, strand asymmetry and sequence orientation.

Analysis of the upstream region

The differences between upstream sequence that occurs close to the TSS and sequence that is more distant were studied. Sequence trends across the 5' upstream region in different positional locations are likely to reflect differences in the organization of sequence and therefore relate to structural and functional changes. This would provide a better understanding of how the cell utilizes the upstream region to regulate gene expression.

It has already been mentioned that the 10 Kb upstream sequence possesses regulatory elements and spacer, etc... that are likely organized in specific ways. Also, there are presumably boundaries where these different 'functional units' occur. Therefore changes across the different positional locations of the upstream sequence were analyzed in order to see potential changes and boundaries.

A global analysis of the upstream sequence of numerous human genes was carried out, i.e. across all the known genes. No distinction was made between different gene/promoter types since the intention was to study only general trends and build an average gene picture. Both repeat-containing and repeat-free fractions were utilized. This enabled the effect of repeats to be observed.

Only the sense strand was used for the analysis because gene regions were of interest and the information being obtained was of a directional nature, since transcription occurs in the downstream orientation. Also, strand bias may important for transcription and protein binding to the DNA. This is generally in line with the method used by Louie et al, 2003, whereby only one strand was used for the upstream sequence analysis of mononucleotides and other sequence properties. Sequence was therefore taken in the 5' to 3' orientation.

The upstream sequences were also subjected to filtering so as to refine the dataset. This involved eliminating the upstream sequences of mRNAs derived from hypothetical proteins or predicted mRNAs. There are clearly arguments for and against this type of filtering that was designed to remove predicted or unverified gene sequences. On the one hand there is a loss of data whilst on the other hand a more accurate or refined dataset was yielded.

For all measurements average values were taken across the entire dataset of human genes. This averaging served to increase the signal relative to noise and helped to display an overall picture or trend across this 10 Kb upstream sequence, i.e. providing an 'average gene' result.

The 10 Kb sequence upstream of the TSS was analysed for dinucleotide composition as ten separate 1 Kb sliding-window segments. Also, the 2 Kb upstream of the TSS was analysed as 250base segments. An important issue is whether the use of such small sequence segments is accurate enough or appropriate for a dinucleotide analysis. The dinucleotide profile may be noisy when small sequence segments are used. In fact in past experiments it was mostly large sequence stretches (50 Kb) of genomic regions that were used as samples (Karlin et al, 1997, Burge et al, 1992). However, experiments by Jernigan et al, 2002, showed that dinucleotide relative abundance profiles and the genomic signature are stable at much smaller intervals, even as small as 125 bases.

Dinucleotide composition and the dinucleotide odds ratio

The dinucleotide representation is a value that can be used to assess dinucleotide contrasts whilst taking into consideration the mononucleotide composition of the sequence. This describes the proportion of each dinucleotide, above or below the random expectation. Dinucleotide representation was calculated by using an odds ratio. The odds ratio can also be referred to as the single-strand dinucleotide relative abundance ratio (Karlin et al, 1995).

Odds ratio: $\rho xy = fxy / fxfy$

fx is the frequency of the nucleotide X within the sequence and fxy is the frequency of the dinucleotide XpY within the sequence. The result obtained from a frequency /count of nucleotides (and dinucleotides) is then multiplied by n (where n = length of sequence) in order to standardise the odds ratio. Alternatively the odds ratio may be calculated directly from nucleotide (and dinucleotide) proportions. $\rho xy \gg 1$ indicates over-representation of the dinucleotides, whereas $\rho xy \ll 1$ indicates under-representation. In a random sequence (i.e. a shuffled sequence) the ρxy values for all the dinucleotides approach 1.0.

The odds ratios of the sixteen dinucleotides form dinucleotide relative abundance profiles, whose difference from 1, provide a measure of deviation from randomness. It has been determined (Karlin et al, 1998) that for a random sequence the **ρxy** values have the following relationship; the deviation from 1 is approximately $\mathbf{1/\sqrt{n}}$. For $\mathbf{n \sim 1000}$, $\mathbf{|\rho xy - 1| = 0.031}$.

Dinucleotide relative abundance profiles are highly stable for bulk DNA. It is thought that the reason for this may be the existence of genome-wide factors. Examples include the replication and repair machinery, mutational tendencies and structural tendencies of genomic DNA. The dinucleotide relative abundance profiles show a departure from

randomness of genomic DNA sequences and collectively form a genomic signature. Therefore this has been used a means to study compositional differences between organisms (Karlin et al, 1995). Compositional differences have also been analysed in this way within organisms (Karlin et al, 1997), such as the comparison between mitochondrial and nuclear genomes. Also similar studies were carried out (Gentles et al, 2001) for human genome chromosomes 21 and 22 bulk DNA. Here it was found that the difference between these chromosomes was similar to the differences within the chromosomes.

Comparisons have been made (Karlin et al, 1994) for relative abundance differences between di-, tri- and tetra-nucleotides. There is actually a high correlation between these. This means that DNA structural/conformational arrangements are primarily dependent upon dinucleotide steps (Breslauer et al, 1986). This is one reason for the use of dinucleotides and not tri- and tetra- nucleotides in this study. Also, as already explained, the structural tendencies of dinucleotide steps have been well characterised. Compositional changes across the 10 Kb upstream sequence were analysed by using dinucleotide compositions and the odds ratio. The aim was to assess overall structural tendencies and their changes across the upstream sequence of the human gene.

Measuring strand asymmetry and sequence orientation

In theory if nucleotide substitutions occur symmetrically in both DNA strands the probability of a nucleotide transition event would have a strand symmetric relationship because of Watson-Crick base pairing. If in one strand the nucleotide A is substituted for G with probability P_{AG} and in the other strand with probability Q_{AG}, the following relationship would be true if there is in fact strand symmetry; $P_{AG} = Q_{AG}$, $P_{AG} = P_{TC}$ and $Q_{AG} = Q_{TC}$. This relationship may also be extended to dinucleotides.

The aim of this experiment was to extend the work carried out by Louie et al, 2003 in analysing strand asymmetry in the upstream region of the human gene. However, here the analysis is extended to one of dinucleotides. If there is strand symmetry, the frequency of the following dinucleotides is expected; ApA = TpT, ApC = GpT, ApG = CpT, TpC = GpA, TpG = CpA, CpC = GpG. In this experiment these dinucleotides are referred to as asymmetric pairs.

Strand asymmetry may be measured in the DNA sequence, within any given strand in terms of (C-G)/(C+G) and also (A-T)/(A+T), (Mrazek et al, 1998, Lobry, 1996). This

describes the skew of mononucleotide frequencies in the sequence. For dinucleotides this frequency skew is also given by (Shioiri et al, 2001):

Skew of dinucleotide frequencies for asymmetric pairs:

(fxy – fx'y') / (fxy + fx'y')

This is similar to the mononucleotide index. In this expression, X'Y' are symbols for the inverted complementary dinucleotide of XpY. The dinucleotides odds ratio that accounts for strand asymmetry in the DNA sequences is given by (Burge et al, 1992):

Skew of dinucleotide odds ratios for asymmetric pairs:

2(fxy + fx'y') / ((fx + fy') (fy + fx'))

There are six dinucleotide directional pairs; ApT and TpA, ApC and CpA, TpC and CpT, TpG and GpT, CpG and GpC, ApG and GpA. This set of directional pairs together, and the skew in their frequencies in a single strand describe a tendency within the DNA sequence for the presence of one dinucleotide in the pair over the other. The expressions for skew in dinucleotide frequencies and odds ratios are as follows: In these expressions fxy refers to the frequency of the dinucleotide XpY and fyx to the frequency of its directional opposite YpX.

Skew of dinucleotide frequencies for directional pairs:

(fxy – fyx) / (fxy + fyx)

Skew of dinucleotide odds ratios for directional pairs:

(fxy + fyx) / 2 (fx fy)

The aim was to determine changes in strand asymmetry and sequence orientation specifically with respect to dinucleotides across the 10 Kb upstream sequence. This was carried out utilising the expressions given above.

2.2 Methods

2.2.1 Obtaining sequences from the human genome database

The DNA sequences for this project were obtained from the NCBI human genome database, build 35. The primary region of interest was the 10 Kb 5' upstream sequence of protein coding genes. As well as this region, sequences from the first exon, the first intron and also the genome-wide sequence were used in the analysis. By the term genome-wide is meant the sequence obtained from the entire human genomic DNA content, i.e. from all of the contigs.

The upstream sequence dataset

The upstream sequence of the gene was taken to begin at the TSS; just one nucleotide upstream of it. This sequence therefore excluded the proposed start site. 10 Kb of the 5' upstream sequence of each gene was taken utilizing the contig annotations from the set of human genome NCBI files. The resultant dataset was then filtered in order to remove duplication and any poor quality data, thereby ensuring a maximum level of accuracy. The location of each mRNA sequence was checked against the human genome contig. utilizing the annotations file. Following this, an all-against-all search for these mRNA locations was made. This means that an mRNA location was checked against all other mRNAs. If there was any overlap of the mRNA or its upstream sequence with that of another mRNA, one of the mRNA's would be eliminated from the final dataset. Also, the annotations were utilized to determine whether the mRNA sequence was on the forward or reverse strand and the upstream sequence was extracted accordingly. The obtaining of the mRNA dataset was carried out using the program *Upstream_LOCATE* (See appendix E.1 for more information).

The sequences were then subjected to a further level of filtering so as to refine the final dataset. This involved eliminating the sequences that were derived from hypothetical proteins or predicted mRNA's. The total number of mRNA sequences was 28,162 from the build 35 human genome NCBI file; rna.fa. Word searches were carried out on the annotations for these mRNA sequences and any sequences that were hypothetical were removed from the final dataset. See appendix E.2 for more details.

The number of mRNA sequences remaining post-filtering of the annotations file was 18,832. Additional sequences were then eliminated due to poor quality (runs of n's or unidentified bases), leaving a final dataset of 18,725 mRNA's. For the final dataset of (post-

filtering) 18,725 mRNA's, 10 Kb of upstream sequence was extracted. This 10 Kb sequence was then sub-divided into ten 1 Kb non-overlapping portions. From the 5' to the 3' end, these were labeled; *upstream10, upstream9, upstream8, upstream7, upstream6*.....and *upstream1*. The dataset named *upstream1* is therefore closest and just adjacent to the TSS. A second upstream dataset was also compiled. This was 2 Kb in length just upstream of the TSS and was subdivided into eight 250 base sequence portions. This second dataset was formed since the most prominent dinucleotide changes occurred over the 2 Kb region upstream of the TSS.

The first exon, first intron and the genome-wide sequence

Samples of the first intron and the first exon of genes were extracted from the genome database. This was done using the NCBI annotations files for each of the contigs as was carried out for the upstream sequences. From the dataset of exon1 sequences another dataset was then derived which, comprised of only the coding sequence excluding the 5' UTR. Intron and exon sequences less than 100 bases in length were excluded.

Genome-wide sequence was also taken for the analyses. Genomic DNA sequence was from all contigs of all the chromosomes from NCBI human genome build 35. These were sampled by taking sequence from only one strand of the strands, the one given in the contig file. This of course does not correspond to any particular chromosomal strand, so that no distinction has been made between chromosomal strands.

The repeat masked dataset

Datasets of repeat masked upstream sequences were generated for each of the ten 1 Kb upstream sequence segments (*upstream1-to-upstream10*) described above. The NCBI masked human genome sequence files for build 35 were utilized. These are human genome sequences that have been masked for all known human repeats by Repeat Masker (*www.repeatmasker.org*)

From this pre-masked genomic DNA sequence, the 10 Kb upstream sequence was extracted exactly as was done for the unmasked dataset (above). I.e. the identical set of 18,725 mRNAs was utilized and the upstream sequence taken for this dataset from the masked genome sequence files. The 10 Kb sequence was then sub-divided into ten 1 Kb sequence portions as before. This time though if more than 90% of each of the 1 Kb upstream sequence fragments was masked, that particular fragment was eliminated from the final dataset.

2.2.2 Dinucleotide Composition

Dinucleotide proportions were taken for each of the ten 1 Kb upstream datasets; *upstream1-to-upstream10,* and also for intron1, exon1, coding1 and the entire genomic DNA sequence. There are a total of sixteen different possible dinucleotides; ApA, ApT, ApC, ApG, TpA, TpT, TpC, TpG, CpA, CpT, CpC, CpG, GpA, GpT, GpC and GpG, Each of these dinucleotides was determined stepwise along the sequence from the 5' to the 3' along the transcribed strand.

For any given sequence dataset, such as *upstream1*, an average (median) dinucleotide proportion measure was taken for each individual dinucleotide across the entire dataset of 18,725 DNA fragments. This averaging for each dinucleotide proportion across the entire dataset was carried out for each of the ten upstream datasets, for *exon1, coding1* and *intron1*. For the genome-wide sequence, each dinucleotide proportion was worked out for each human chromosome individually and the result for that dinucleotide was then averaged across all the chromosomes.

T-tests (two-tailed at the 5% level of significance) were carried out for each dinucleotide proportion in adjacent location sets, i.e. *upstream1* compared with *upstream2, upstream2* compared with *upstream3* etc... This means for example, that the proportion of a dinucleotide XpY in 18,725 fragments of *upstream1* would be compared with the equivalent in *upstream2*. Each of these datasets contains sequence fragments that are physically proximal to one another. This analysis would determine at which locations along the 10 Kb upstream (divided into 1 Kb portions) changes in dinucleotide composition occur.

2.2.3 Dinucleotide representation

Calculation of the odds ratio

The dinucleotide representation of each of the sixteen dinucleotides was carried out individually for each of the ten upstream datasets; *upstream1-to-upstream10, intron1, exon1* and *coding1* and also for the *whole genome*. This was done using the odds ratio; $\rho xy = fxy/fxfy$. Within each upstream dataset, e.g., *upstream1*, odds ratio values were calculated for each

individual 1 Kb sequence fragment taking into consideration its specific nucleotide and dinucleotide proportions. The odds ratio results were then averaged (using the median) over the entire dataset of 18,725 sequence fragments.

For the genome-wide sequence, the nucleotide and dinucleotide proportions were taken across all the contigs of each individual chromosome. E.g. the frequency of dinucleotide XpY and nucleotides X and Y were found in the entire sequence of chromosome1. The odds ratio could then be worked out. This odds ratio was then found for XpY in all the human chromosomes and an average (mean) value was calculated for the odds raito of XpY across all the chromosomes, thereby giving an odds ratio value for XpY in the 'entire' human genome. A program was written called *DINUC_COMP* that obtained; mononucleotide frequencies, mononucleotide proportions, dinucleotide frequencies, dinucleotide proportions and then calculated the odds ratio for each upstream sequence in the set of 18,725. Also, some basic descriptive statistics were calculated by the same program. See appendix E.3 for more details.

A statistical analysis of real versus random dinucleotide proportions

T-tests were carried out (two-tailed at the 5% level of significance) to determine whether there is a significant difference between the dinucleotide proportions of a real dataset of upstream sequences, such as *upstream1* and its equivalent random (or shuffled) dataset of identical mononucleotide proportions. These random proportions were the theoretical proportions for the random upstream sequence. This was done for each dinucleotide within each upstream segment. For example, The actual proportion of the dinucleotide, XpY in each of the 18,725 sequence fragments of *upstream1* was taken. This was then compared with the theoretical proportion of XpY in the equivalent random set of these 18,725 sequence fragments. These real versus random statistical tests for dinucleotide proportions were carried out for each of the upstream sequence segments; *upstream1-to-upstream10.*

2.2.4 Sequence directionality and strand asymmetry

Skew of dinucleotide frequencies for asymmetric pairs

For the six asymmetric pairs; ApA and TpT, ApC and GpT, ApG and CpT, TpC and GpA, TpG and CpA, CpC and GpG, the skew of dinucleotide frequency was worked out

using the following expression; (fxy-fx'y') / (fxy+fx'y'). This expression was utilized to work out the skew for each of the 18,725 1 Kb sequence fragments for *upstream1*. The result for the skew of dinucleotide frequency was then averaged out (using the median) over this entire dataset of 18,725 sequences. The same procedure was then repeated for *upstream2-to-upstream10*. Statistical tests were then carried out to compare the frequency of each asymmetric dinucleotide pair. These were paired t-tests, two-tailed at the 5% level of significance.

Skew of dinucleotide odds ratios for asymmetric pairs

The skew of odds ratio was then measured for each of the six asymmetric pairs using the following expression; 2(fxy+fx'y') / ((fx+fy')(fy+fx')). This expression was utilized to work out the odds ratio skew for each of the 18,725 1 Kb sequence fragments for *upstream1*. The result was then averaged out (using the median) over this entire dataset of 18,725 sequences. The same procedure was then repeated for *upstream2-to-upstream10*.

Skew of dinucleotide frequencies for directional pairs

For the six directional pairs; ApT and TpA, ApC and CpA, TpC and CpT, TpG and GpT, CpG and GpC, ApG and GpA, the skew of dinucleotide frequency was worked out using the following expression; (fxy-fyx) / (fxy+fyx). This expression was utilized to work out the skew for each of the 18,725 1 Kb sequence fragments for *upstream1*. The result for the skew of dinucleotide frequency was then averaged out (using the median) over this entire dataset of 18,725 sequences. The same procedure was then repeated for *upstream2-to-upstream10*.

Statistical tests were then carried out to compare the frequency of each dinucleotide pair. These were paired t-tests, two-tailed at the 5% level of significance. For instance the frequency of XpY was compared with YpX for the 18,725 1 Kb sequence fragments of *upstream1*. This same test was then repeated for *upstream2-to-upstream10*.

Skew of dinucleotide odds ratios for directional pairs

The skew of odds ratio was then measured for each of the six directional pairs using the following expression; (fxy+fyx) / 2fxfy. This expression was utilized to work out the odds ratio skew for each of the 18,725 1 Kb sequence fragments for *upstream1*. The result was

then averaged out (using the median) over this entire dataset of 18,725 sequences. The same procedure was then repeated for *upstream2-to-upstream10.*

2.2.5 The effect of Repeats

The level of repeat masking and hence the proportion of repeat sequence was worked out for each of the different upstream positional segments; *upstream1-to-upstream10.* First of all the proportion of masked nucleotide was worked out for each of the 18,725 1 Kb sequence fragments of the *upstream1* region individually. This proportion of masked nucleotide was averaged (median) over the entire dataset of 18,725 sequence fragments, giving an average value for extent of repeat masked sequence within the *upstream1* region. This procedure was then repeated for each of the *upstream2-to-upstream10* datasets, enabling the proportion of relative masking across the 10 Kb upstream segments to be seen.

In order to see the effects of repeats on dinucleotide proportion and representation the experiments described above were carried out again, this time utilizing the equivalent repeat masked upstream sequence dataset. However, some important differences must be noted. If in any single 1 Kb upstream fragment more than 90% of the nucleotides were masked for repeats that particular fragment was eliminated. Due to the repeat masking of each of the 18,725 1 Kb sequence fragments, the remaining unmasked number of nucleotides varied for each fragment. Nucleotide composition was therefore measured as a proportion of the total number of unmasked nucleotides.

2.3 Results

2.3.1 Dinucleotide composition

Specific dinucleotides: an increase or decrease in proportion across the upstream

The trends in dinucleotide compositional changes across the upstream can be divided into four general categories (see figure 2.1). These dinucleotide composition changes occur mainly across the 4Kb sequence upstream of the TSS. However, here only the 2Kb sequence is shown because the most prominent changes are in this region. For full 10Kb results and plots see Appendix A.3.

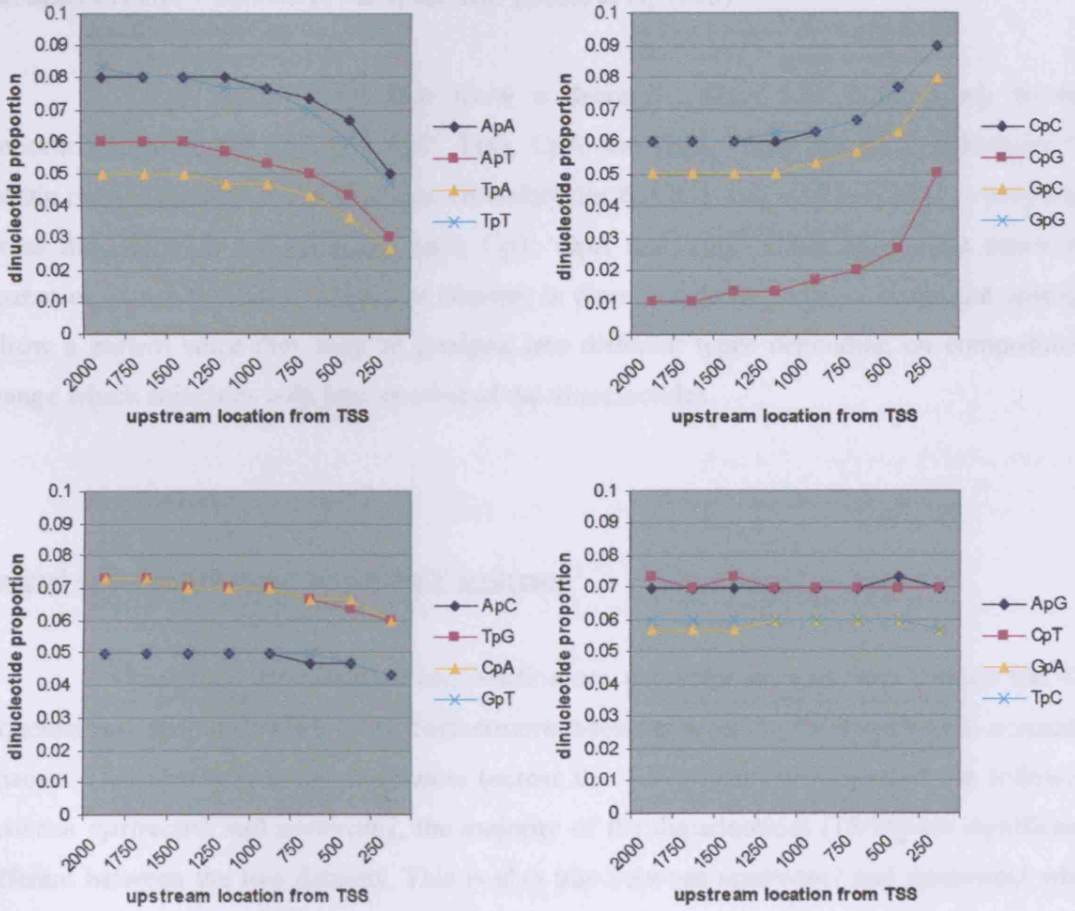


Figure 2.1: Graphs showing changes in dinucleotide proportions across the 2 Kb upstream sequence.

These are divided into four categories depending on the observed change across the upstream sequence and nucleotide content.

There is a gradual change in the proportion of certain dinucleotides. There is a decrease in the proportion of dinucleotides comprising of two weak (hydrogen-bonding) bases; ApA, ApT, TpA and TpT towards the TSS and an increase in dinucleotides with two strong (hydrogen bonding) bases; CpC, CpG, GpC and GpG.
Also, there is a slight decrease in proportion of ApC, TpG, CpA and GpT towards the TSS, consisting of one purine and one pyrimidine base.

The most marked differences in dinucleotide composition across the upstream sequence are seen for the following; ApA, ApT, TpA, TpT, CpC, CpG, GpC and GpG. There is a decrease in the proportion of dinucleotides comprising of two weak (hydrogen bonding) bases towards the TSS and an increase in dinucleotides with two strong (hydrogen bonding) bases. This result is in line with the general increase in C and G mononucleotide content, and a decrease in A and T content towards the TSS (Louie et al, 2003).

Other dinucleotides that show a decrease (albeit less pronounced) in their proportion towards the TSS are; ApC, TpG, CpA and GpT. These are all dinucleotides that contain one purine and one pyrimidine. Dinucleotides that show no visible change in proportion across the upstream sequence are ApG, CpT, GpA and TpC. These all contain either two purines or two pyrimidines. Therefore changes in dinucleotide compositon across the upstream follow a pattern since they may be grouped into different types depending on compositional change which coincides with base content of the dinucleotides.

General dinucleotide trends across the 5' upstream

Changes in dinucleotide composition are generally seen to occur within the 4Kb sequence just upstream of the TSS. Furthermore, t-tests comparing the dinucleotide content of adjacent 1Kb upstream segment datasets (across the 10Kb upstream) revealed the following: Between *upstream1* and *upstream2*, the majority of the dinucleotides (15/16) are significantly different between the two datasets. This is also true between *upstream2* and *upstream3* where 12/16 dinucleotides are significantly different (see table 2.1).

There are some, (although a minority of) dinucleotides that are significantly different between *upstream3* and *upstream4*, only 5/16 dinucleotides. Most of the further upstream datasets show no change in dinucleotide content. A notable exception is the *upstream7* and *upstream8* comparison for ApA and also CpC, for which these datasets were found to be

significantly different. The reason for this is unknown. See also ANOVA results for this data (Appendix A.4).

Dataset pairs / dinucleotides	upstream9/10	upstream8/9	upstream7/8	upstream6/7	upstream5/6	upstream4/5	upstream3/4	upstream2/3	upstream1/2
ApA	0.6503	0.1889	0.0176	0.9294	0.9265	0.0409	0.4690	0.0000	0.0000
ApT	0.7083	0.2995	0.2474	0.8341	0.6237	0.5772	0.0002	0.0000	0.0000
ApC	0.0976	0.6111	0.9255	0.6585	0.1285	0.9203	0.9768	0.5980	0.0000
ApG	0.9259	0.3326	0.8987	0.7919	0.8385	0.7894	0.8619	0.0822	0.0007
TpA	0.5010	0.2411	0.3513	0.8788	0.6384	0.8979	0.0158	0.0000	0.0000
TpT	0.3158	0.8087	0.9240	0.8508	0.4598	0.8163	0.7897	0.0000	0.0000
TpC	0.8512	0.3910	0.0524	0.5937	0.9056	0.2671	0.5334	0.9836	0.0002
TpG	0.7675	0.9189	0.1544	0.4756	0.4950	0.9061	0.3351	0.0000	0.0000
CpA	0.0532	0.4493	0.5556	0.9806	0.3686	0.6098	0.0108	0.0000	0.0000
CpT	0.7636	0.7894	0.0836	0.8589	0.5233	0.4140	0.0614	0.0203	0.0000
CpC	0.8260	0.8194	0.0369	0.6516	0.5355	0.1557	0.1278	0.0000	0.0000
CpG	0.9997	0.4747	0.1861	0.9275	0.8572	0.7439	0.0001	0.0000	0.0000
GpA	0.9191	0.8898	0.3140	0.9896	0.8503	0.5129	0.8155	0.0181	0.1674
GpT	0.7225	0.6018	0.2345	0.7094	0.9198	0.1412	0.9199	0.1314	0.0000
GpC	0.9575	0.3348	0.1747	0.6382	0.3852	0.8020	0.0202	0.0000	0.0000
GpG	0.9938	0.6212	0.5195	0.7075	0.9505	0.2384	0.1195	0.0000	0.0000

Table 2.1: Statistical Testing: Comparison of dinucleotide composition of adjacent upstream segments:

The following is a summary table of the dinucleotide t-test P-values of adjacent upstream segments, each 1Kb apart. Nine dataset comparisons are shown across the 10Kb sequence. The result for each pair of adjacent upstream datasets is shown, for instance, *upstream1* is compared with *upstream2*, *upstream2* with *upstream3*, etc... for each of the sixteen dinucleotides. P values highlighted in red show those datasets pairs found to be significantly different for a particular dinucleotide.

This table also provides a visual depiction of changes in dinucleotide proportions that occur across the 10Kb upstream.

The result above suggests a potential boundary at around the 3-4Kb upstream location regarding dinucleotide composition changes. Beyond the 4 Kb region upstream of the TSS, very few changes in dinucleotide composition occur, with respect to the 1Kb sequence segments. The result also shows that boundaries of dinucleotide change vary depending on the particular dinucleotide and this probably relates to structural features which will be discussed.

2.3.2 Dinucleotide representation

In general, across the upstream sequence, if a particular dinucleotide is over-represented it remains over-represented throughout the sequence (see figure 2.2). The changes across the upstream are not so dramatic that a huge shift occurs from under- to over-representation.

The following dinucleotides are under-represented; CpG, TpA, ApC, GpT, and ApT. The following dinucleotides are over-represented; ApA, ApG, TpT, TpG, CpA, CpT, CpC, GpG. Under-represented dinucleotides are those that have been specifically suppressed since they are present at a proportion that is lower than expected given the base composition of the sequence. The opposite is true for over-represented dinucleotides.

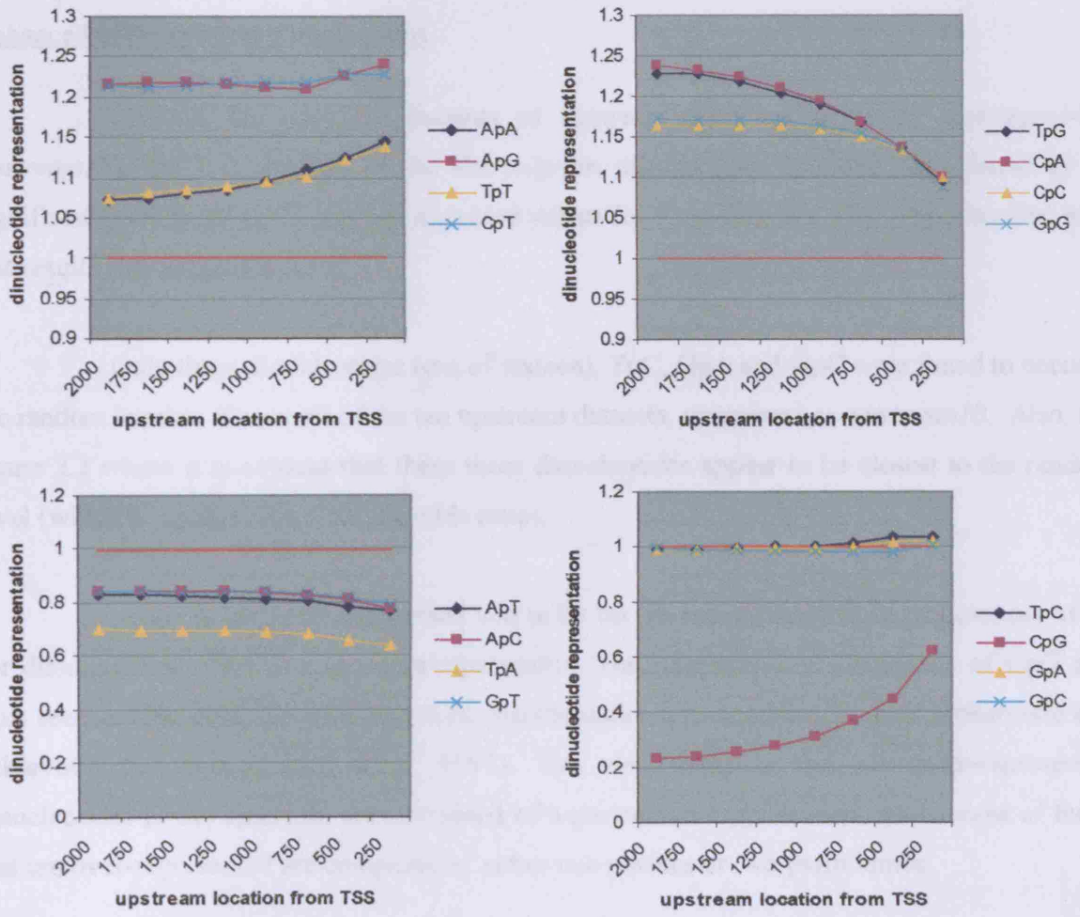


Figure 2.2: Graphs showing changes in dinucleotide representation (odds ratio: ρxy) across the 2Kb upstream sequence.

A distance from randomness value above 1.0 implies an over-represented or enhanced dinucleotide; a value below 1.0 indicates an under-represented or suppressed dinucleotide. A value of 1.0 means a dinucleotide is present at the randomly expected level. The random expectation is shown as a red line.

There are four general different possible changes that may occur across the sequence regarding dinucleotide representation, hence the division of results into four graphs.

Of the dinucleotides that are under-represented some become more suppressed towards the TSS. These include ApT, ApC, TpA and GpT, all of which contain one purine base and one pyrimidine. Others may become less suppressed in the sequence, in this case only CpG.

Of those dinucleotides that are over-represented, some become more enhanced towards the TSS; ApA, TpT, ApG, CpT. These contain either two purine or two pyrimidine bases. Others may become less enhanced; TpG, CpA, CpC, and GpG.

Enhanced or suppressed dinucleotides

Within the ten 1Kb datasets of upstream sequence segments (*upstream1*-to-*upstream10*), the composition of the above-given thirteen dinucleotides were found to be significantly different to the random expected values for these datasets. This was according to t-test results (see appendix A.6).

Only three dinucleotides (out of sixteen), TpC, GpA and GpC were found to occur at the random level in almost all of the ten upstream datasets, *upstream1*-to-*upstream10*. Also, see figure 2.2 where it is evident that these three dinucleotides appear to be closest to the random level (which is a value of 1.0 for the odds ratio).

CpG is the most suppressed and is by far the most distant from randomness of all the dinucleotides. TpA is also under-represented. The relatively low abundance of CpG and TpA seen in this work fits well with their general under-representation in most prokaryotic and eukaryotic genomes (Karlin et al, 1997). The main issue is that all under-represented dinucleotides in the upstream are composed of a purine and a pyrimidine, whilst most of those that are over-represented are composed of either two purines or two pyrimidines.

Changes in dinucleotide representation across the upstream sequence

Superimposed on the general baseline of dinucleotide over- or under-representation were some changes across the 10Kb upstream, with some dinucleotides becoming more (or less) over-represented and others becoming more (or less) under-represented. These changes were as follows. The dinucleotides; ApT, ApC, TpA and GpT become more under-represented towards the TSS. Interestingly these are each composed of one purine base and one pyrimidine base. In contrast to this, ApA, ApG, TpT and CpT become more over-represented. These are all composed of either purine only or pyrimidine only bases. All of the above dinucleotides are therefore ones that become more distant from the random expectation in the direction of the TSS. These observations and this specific shift imply that there is an importance in the changed representation of purines and pyrimidines across the 5' upstream.

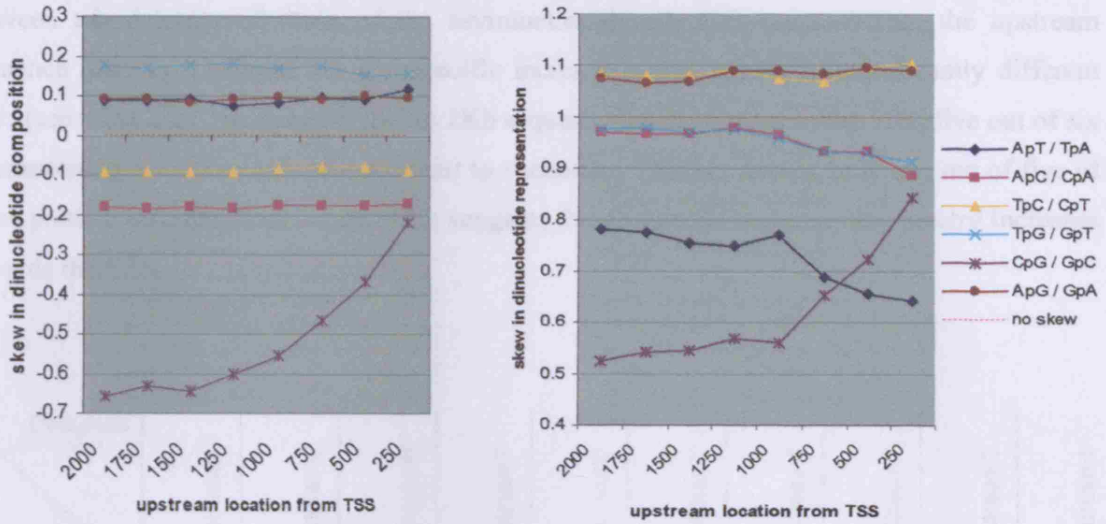
The following are dinucleotides that become closer to the random model towards the TSS. CpG becomes less under-represented. TpG, CpA, CpC, and GpG become less over-represented. The steepest difference across the 2Kb upstream region is for CpG, then CpA and TpG for the odds ratio. This means that these dinucleotides display the most dramatic change in representation across the 2Kb upstream sequence (this is also true for the 10Kb upstream sequence).

2.3.3 Sequence directionality and strand asymmetry

Skew in directional pair composition and representation

All dinucleotide directional pairs were found to be signifianctly different (t-tests 5% level) across all the 10Kb upstream datasets, with respect to their composition. Changes in the directional pairs occur over the 5Kb sequence upstream of the TSS.

Whilst there is skew in composition between all the dinucleotide directional pairs (of at least 10%), the actual level of this skew does not change much across the sequence for most of these pairs. See figure 2.3 for changes across the 2Kb upstream sequence. The notable exception is the CpG/GpC pair which possesses a high level skew (1.5-2Kb upstream the skew is around 65%) which is greatly diminished towards the TSS to around 20%). This means that in general sequence directionality decreases towards the TSS.



2.3a. **2.3b.**

Figure 2.3: Graphs of skew in sequence directionality across the 2Kb upstream sequence.
2.3a. Dinucleotide composition is shown.
Directionality exists throughout the 2Kb sequence for all the dinucleotides since there is a skew in composition (non-zero values) of at least approximately +/-0.1 (10%). However for most of the directional pairs this directionality does not change over the 2Kb sequnece. The only exception is the CpG/GpC pair.
2.3b. Dinucleotide representation (the odds ratio) is shown. All the dinucleotide pairs possess a skew in their representation across the upstream. Changes in the level of skew occur for only some of the dinucleotide pairs across the upstream sequence.

For these same directional pairs there is a skew in odds ratio values throughout the upstream. TpC/CpT and ApG/GpA (which are complementary) show little change in this skew across the upstream. ApC/CpA and TpG/GpT (which are complementary) show slight increased skew in the 1Kb sequence towards the TSS. The most prominent changes though across the upstream occur for CpG/GpC for which there is a much reduced level of skew towards the TSS. In contrast, ApT/TpA has an increased level of odds ratio skew in the TSS direction.

Skew in asymmetric pair composition and representation

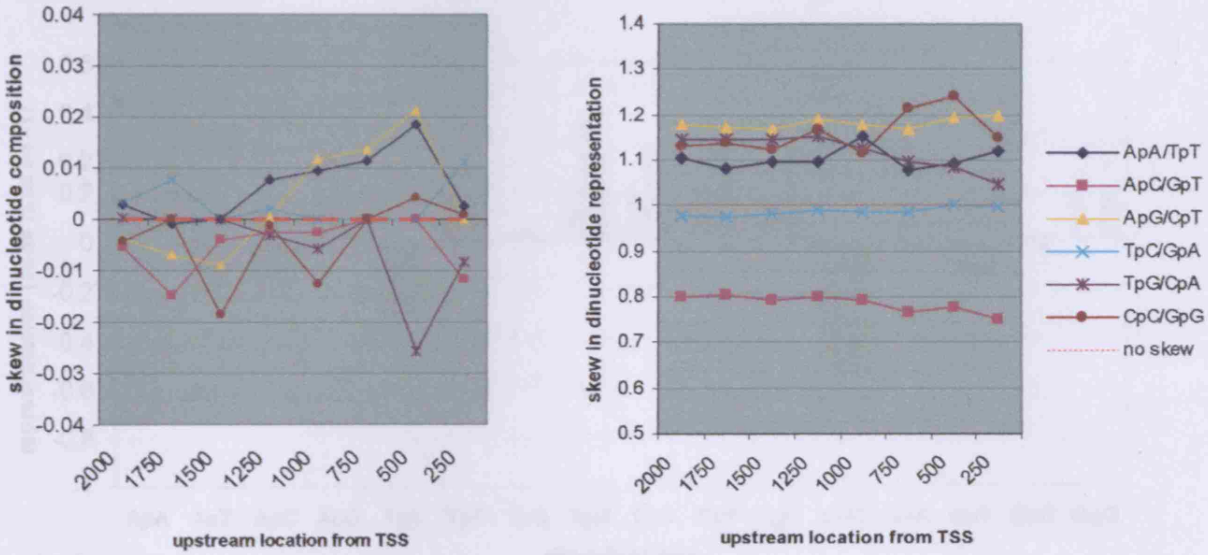
Results show that there is a signifiant difference (t-tests 5% level of signifiance) between the datasets of some of the asymmetric dinucleotide pairs. Across the upstream sequence there is a general but non-specific increase in the number of significantly different pairs (see table 2.2). For example, in the 1Kb sequence dataset closest to the TSS; five out of six asymmetric pairs are signifiantly different to eachother. 10Kb upstream, only one out of five of these pairs is signifiantly different. This suggests that in general sequence asymmetry increases towards the TSS.

Datasets / Asymmetric pairs	upstream10	upstream9	upstream8	upstream7	upstream6	upstream5	upstream4	upstream3	upstream2	upstream1
ApA/TpT	0.932	0.165	0.003	0.641	0.451	0.916	0.070	0.019	0.101	0.000
ApC/GpT	0.083	0.957	0.415	0.056	0.049	0.509	0.493	0.552	0.807	0.001
ApG/CpT	0.870	0.607	0.922	0.081	0.094	0.038	0.002	0.000	0.333	0.000
TpC/GpA	0.301	0.330	0.117	0.405	0.211	0.131	0.003	0.001	0.180	0.004
TpG/CpA	0.011	0.571	0.248	0.514	0.191	0.922	0.501	0.034	0.601	0.427
CpC/GpG	0.247	0.397	0.632	0.104	0.653	0.733	0.980	0.990	0.093	0.024

Table 2.2: Summary table of significance tests for difference between asymmetric dinucleotide pairs.
Results are given for the 1Kb datasets spanning 10Kb of sequence upstream of the TSS. t-test P values are shown and highlighted in red for asymmetric pairs that are significantly different at a particular upstream location.
This table shows that there is a general increase in strand asymmetry in the 10Kb sequence towards the TSS.

The result for changes in skew of asymmetric dinucleotide pairs is less clear. Here the difference between asymmetric pair composition is variable across the upstream sequence (see figure 2.4a). Although in general there appears to be an increase in the difference between asymmetric pairs towards the TSS. This is true both for the 10Kb upstream sequence and 2Kb sequence.

Most of the aymmetric pairs display a skew in their odds ratio values of around between 10-20%, depending on the dinucleotide pair (see figure 2.4b). TpC/GpA is the only exception since it possesses very little skew. However, there appears to be no clear change across the upstream in this value for all the dinucleotide asymmetric pairs. Hence there appears to be some increase in strand asymmetry towards the TSS with respect to dinucleotide compostion. However, whilst there is dinucleotide representation strand asymmetry, it does not appear to change across the upstream.



2.4a. **2.4b.**

Figure 2.4: Graphs showing changes in strand asymmetry across the 2Kb upstream sequence.

2.4a. This shows the skew in asymmetric dinucleotide pair composition. Whilst there is a wide variation and fluctuation in this value across the sequence, there appears to be a general increase in skew of these asymmetric paris towards the TSS.

2.4b. This shows the skew in the odds ratio for asymmetric dinucleotide pairs.

2.3.4 Comparison of the upstream with other genomic regions

In general dinucleotides that are under-represented are under-represented in all the different genomic regions (see figure 2.5). The same is true regarding over-representation.

These profiles are likely the result of the necessity for stable DNA helical structure. Also in general (but not always) the dinucleotide distance from profile for the non-coding sequences is similar (i.e. upstream and intronic) whereas the *exon1* and *coding1* sequences are very different to the non-coding sequences. This is to be expected since similar sequence types are likely to have similar structural features. This may be because superimposed upon a baseline structural requirement for genomic DNA in general there is tendency for specific sequence types to possess their own characteristic sequence motifs related to their particular function.

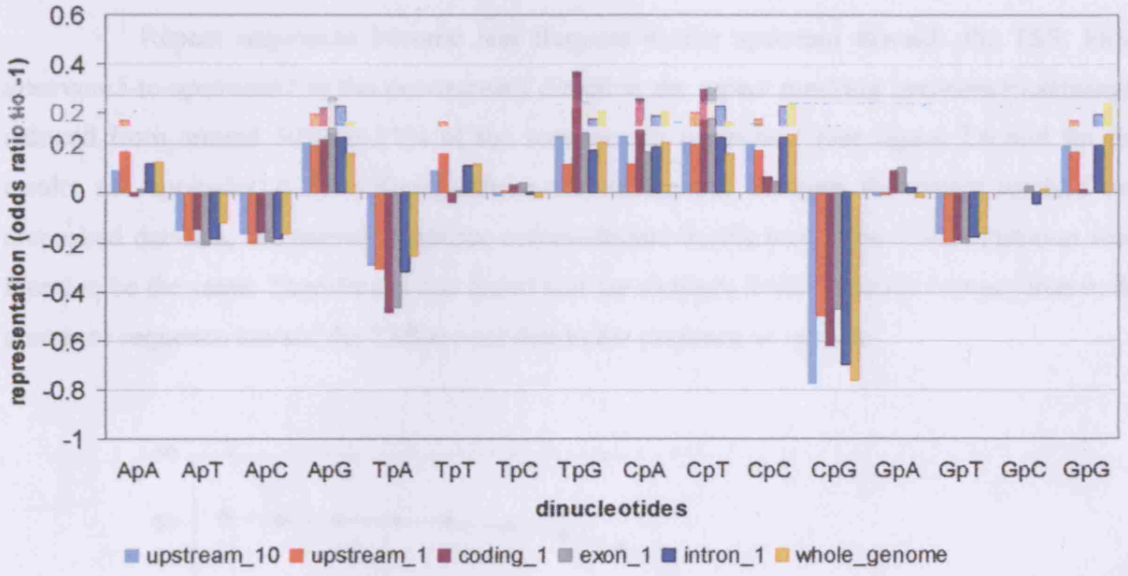


Figure 2.5: The representation of each dinucleotide (odds ratio: ρxy -1) in different genomic regions.

The first exon (*exon1*), the coding sequence of exon1 (*coding1*), the first intron (*intron1*) and *upstream1* and *upstream10* and the *whole genome* sequence are shown. A representation value of zero is the random expectation. A value above zero indicated over-representation and a value below zero under-representation.

The most under-represented dinucleotide in the three-sequence types is CpG. Other under-represented dinucleotides include: TpA, ApT, ApC, GpT. The following dinucleotides are over-represented in all three-sequence types: ApG, TpG, CpA, CpT, CpC.

These results show that whilst there is a wide-spread tendency for specific dinucleotides to be either over-/under-represented, the different types of genomic DNA

sequence possess variation in representation values for the individual dinucleotides. The structural and functional differences of the different genomic regions are likely reflected in varying dinucleotide sequence properties.

2.3.5 The effect of repeats

Changes in the extent of repeat masking across the 10Kb upstream sequence

Repeat sequences become less frequent in the upstream towards the TSS. From *upstream5*-to-*upstream1* in the downstream direction the repeat masking becomes increasingly reduced from around 50% to 25% of the sequence in *upstream1* (see figure 2.6 and for full results see appendix A.7). Although there are differences between the repeat masked and unmasked datasets, the overall sequence compositional trends across the 10Kb upstream were found to be the same. Therefore it was found that the changes in dinucleotide composition in the upstream sequence toward the TSS are not due to the presence of repeats.

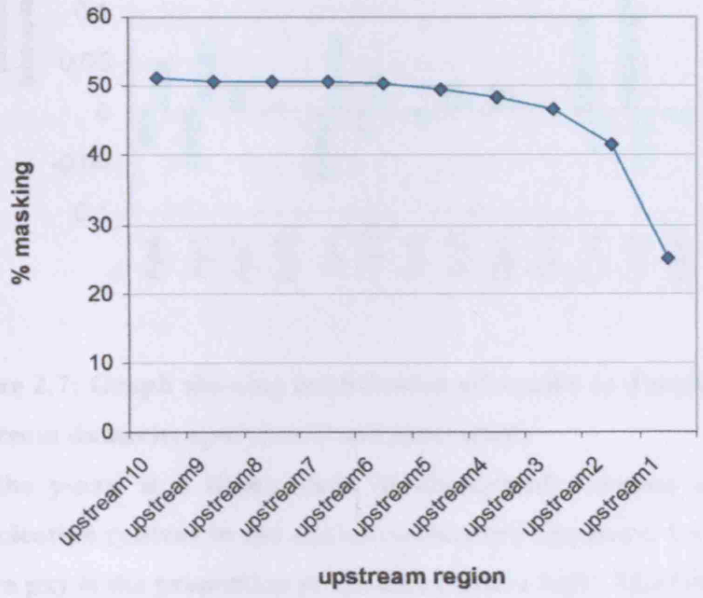


Figure 2.6: The extent of repeat masking across the 10Kb upstream sequence.
Graph of percentage (average) of repeat masked nucleotides for each of the 1Kb positional segments; *upstream1*-to-*upstream10*. Repeat masking is greatly reduced in the TSS

direction. From *upstream5*-to-*upstream1* in the downstream direction the extent of the masking is reduced increasingly from over 50% to 25%.

The influence of repeat sequences on dinucleotide content varies depending on the location of the upstream sequence. See figure 2.7 for ratios of unmasked to repeat masked sequences of dinucleotide composition in; the 1Kb sequences immediately upstream of the TSS (*upstream1*) and also 10Kb upstream (*upstream10*) of the TSS. The most dramatic difference in dinucleotide composition was CpG in *upstream10*. The repeats in this region cause an increase in the CpG content by over 40% (in the unmasked sequence the proportion CpG was 0.0110 and in the repeat masked sequence it is 0.0078).

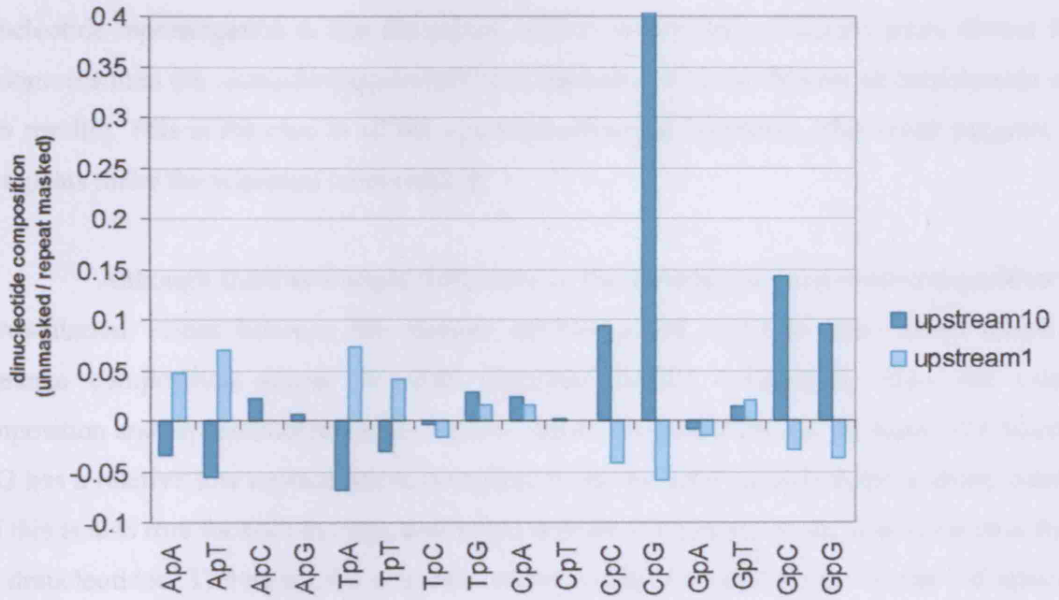


Figure 2.7: Graph showing contribution of repeats to dinucleotide content in two extreme upstream datasets; *upstream10* and *upstream1*.

On the y-axis is a shown ratio of dinucleotide content in unmasked sequence to the dinucleotide content in the equivalent masked sequence. I.e. pxy(unmasked)/pxy(masked) where pxy is the proportion of the dinucleotide XpY. The results reveal the contribution of repeats to the overall dinucleotide content of the sequence. A positive value indicates a positive contribution of the repeat sequences for that particular dinucleotide. For example, in *upstream10* repeat sequences elevate the CpG content by approximately 0.4 (40%). A negative value shows that the repeat sequences diminish the content of that particular dinucleotide.

In fact a marked contrast may be observed between the contribution of repeats within *upstream1* and *upstream10*. In *upstream10* the repeats elevate CpG, GpC, GpG and CpC levels, the dinucleotides containing two strong bases and diminish TpA, ApT, ApA and TpT, the dinucleotides containing two weak bases. For *upstream1* the opposite is true. The repeats cause an elevation of the dinucleotide containing two weak bases and a diminishing of dinucleotides with two strong bases. Therefore the composition of the repeat fraction likely changes over the 10Kb sequence. Furthermore the repeats counter or oppose the change in sequence (dinucleotide) composition across the upstream towards the TSS. This result has also been confirmed by mononucleotide content of the same regions for unmasked and repeat masked sequences (see appendix A.8 and appendix A.9).

The difference between masked and unmasked sequences with respect to dinucleotide representation is that the repeat masked sequences are usually more distant from randomness than the unmasked equivalent (see appendix A.12 for full set of dinucleotide odds ratio results). This is the case in all the upstream positional segments. This result suggests that the repeats make the sequence more random.

Although there is a slight difference in the dinucleotide sequence composition and representation values between the masked and unmasked sequence, the overall trends for sequence composition across the 10kb upstream remain unchanged. Also, the relative composition and representation values between the dinucleotides remain the same. For example, CpG has a relative low representation compared to all the other dinucleotides in every location and this is also true for both the repeat -masked and unmasked sequences. This is the case for all the dinucleotides. The trends for sequence directionality and asymmetry across the upstream also remain the same for the masked and unmasked sequences.

2.4 Conclusions & Discussion

2.4.1 Structural implications

Changes in dinucleotide composition across the upstream sequence imply structural differences and changes in representation indicate a specific (non-random) tendency for certain structures. What do the observed dinucleotide changes across the upstream actually mean in terms of helix conformation and its role in the cell?

The observed suppression of RpY and YpR (dinucleotides containing a purine and a pyrimidine) in the entire 10Kb upstream sequence means that flexible DNA is generally avoided and reduced stability is avoided. This suppression of flexibility becomes even more accentuated towards the TSS. The reason may be a requirement for even greater precision or specificity of structure, i.e. without the multiple possible helical forms afforded by flexible DNA. Since this region (towards the TSS) is read by the transcription machinery, a role which involves specific protein recognition and binding, high structural flexibility may be disadvantageous. In contrast, there may be an advantage for the further upstream intergenic sequence to possess a relatively higher degree of flexibility. This may for example allow for the DNA to be more readily formed into chromatin (Calladine et al, 1986).

It is possible to conclude from the results, that the bistable step SpS (a dinucleotide containing two 'strong' hydrogen bonding bases) is generally suppressed in the upstream. This suppression is alleviated in the TSS direction and SpS composition is greatly increased. Why would bistability be relatively increased towards the TSS? What role could this play in gene regulation? In order to understand this bistability, the situation must be considered in conjunction with other known characteristics. Namely, that flexibility of the DNA is generally more suppressed, stiffness enhanced, and its bistability less suppressed (or relatively increased) towards the TSS.

It may be that flexible DNA allows for an increased propensity for certain types of protein-DNA interactions (Feuerstein et al, 1990). For example, it is thought that flexible DNA can more easily bind histones. Reducing the flexibility of DNA and introducing more rigid DNA may reduce this type of interaction. However, the observed results show that towards the TSS flexibility is further suppressed and rigidity further enhanced. Yet the protein-DNA interaction must still occur.

If the DNA were more rigid though how would the proteins involved in gene regulation bind to it? The answer to this may be the increased bistability that has been observed in this experiment. In this case the DNA still possesses manoeuvrability so that it can adjust itself to the protein (by altering its conformation in order to bind to it) but this adjustment may be far more limited and does not include multiple intermediary conformations as is characteristic of flexible DNA. Therefore this structural characteristic of the DNA may be more specifically discriminatory for regulatory protein binding.

The double helix has the potential to curve in three-dimensions, a feature that is important for its interaction with proteins. Roll angles can determine the potential of the helix to bend in this way. The overall level of curvature though depends on the sum and nature the roll angles of each of the DNA steps along the length of a particular stretch. Of course, there are many different possibilities here.

One example would be interspersed very high roll steps (SpS) with low roll steps (ApA), the high roll ones occurring at every ten bases. This would potentially produce a curvature of around forty-five degrees per turn. This type of sequence is formed mostly of rigid steps. However, since SpS is able to adopt the wide ranging extreme (either high or low) roll angles this sequence has the capability of curvature despite its rigidity. It is this type of structure that may be important for the regulatory protein interaction with DNA.

Often DNA is seen (in X-ray studies) to bend around the regulatory protein. Whilst general flexibility seems to be avoided in the DNA (and even more so towards the TSS), bistable steps are enhanced. This may serve to favour curvature whilst maintaining rigidity. Research by Schatz et al, 1997, on TBP binding to different elements shows that intrinsic DNA curvature and not just flexibility is important for recognition of the DNA by the protein.

In summary, the results of the this work suggest the following; Since flexible DNA is avoided and stiff structure is enhanced towards the TSS, these sequence characteristics may be necessary for gene regulation and could suppress gene activation. The bistable step adds another dimension to the picture. Perhaps then suppressing flexible DNA reduces general protein binding, whilst enhancing bistable DNA permits specific regulatory protein binding. This structural change would therefore affect the way in which proteins bind to the helix.

Also, it is specifically the purine/pyrimidine property of the sequence that becomes less random towards the TSS. I.e. dinucleotides containing one purine and one pyrimidine become more distant from randomness according to the odds ratio. This feature may be attributed to the importance of purines and pyrimidines in determining DNA structure. It

follows therefore that the structural aspect becomes increasingly important in the TSS direction, likely because of regulatory sequences. It may also be the case that this change in purine/pyrimidine dinucleotide representation is important for maintaining sequence stability whilst permitting for other necessary changes in sequence composition.

In general, across the upstream sequence, if a particular dinucleotide is over-represented it remains over-represented throughout the 10Kb sequence. The changes across the upstream are not so dramatic that a huge shift occurs from under- to over-representation.
The balance of dinucleotides is important for maintaining helical structure. It would seem therefore that while there is some variation across the sequence, there is also an overall maintenance of structural stability. Within this general structure there can be some variation that allows for structural and functional specifications in different locations of the DNA.

2.4.2. Sequence directionality and strand asymmetry

It has been shown that the upstream possesses directional bias throughout the sequence studied. This is with respect to all the dinucleotide directional pairs. All six directional pairs were present at significantly different proportions to each other in the upstream sequence and all possessed a skew in composition.

The decreased directionality property of the upstream sequence towards the TSS seems counter-intuitive. This is because the process of transcription is clearly highly directional and requires protein-DNA binding events that are orientation specific. Therefore the sequence would be expected to become more directionally biased towards the TSS due to the high density of protein binding motifs in this region.

It is clear that XpY and YpX are structurally different. These two motifs would therefore be recognised differently by proteins. This is different to the rotational symmetry of sequences that may occur on the different DNA strands, i.e. strand asymmetry. In contrast, differences between the directional pairs (XpY and YpX) in a sequence imply biases in sequence assemly and/or mutation.

General (or baseline) differences in the representation of directional pairs may be due in part to tendancies of DNA sequence formation or assembly that go beyond the initial 'nucleotide mix' or sequence composition. This may be explaied as a specific 5' or 3' nearest neighbour preference.

Superimosed upon this baseline tendency, may be some mutation events generating a bias. For instance, in the upstream sequence (and in genomic DNA in general) the dinucleotide CpG is present at a lower level and is much more suppressed than GpC. Suppression of CpG is thought to be due to methylation and deamination, which causes the mutation of CpG to either TpG or CpA. TpG and CpA occur at a higher than random expectation (whilst CpG occurs at a lower than expected frequency) in the genomic regions analysed. This particular mutation is very common in the DNA of many organisms. The increase in CpG proportion towards the TSS and its decreased suppression may be due to its requirement for regulation and due to the avoidance of this mutation.

One possible explanation for the under-representation of TpA is its presence as a part of regulatory motifs such as the TATA box. TpA becomes more suppressed in the downstream direction. Its suppression would therefore help to prevent the inappropriate binding of regulatory proteins to the DNA.

The difference in representation between CpG and GpC is by far the highest of all the directional pairs. Changes in directionality across the upstream sequence occur between CpG and GpC, ApC and CpA, TpG and GpT. Therefore the overall reduction in directionality towards the TSS may be due to a relative alleviation of mutation events in this direction, and more specficially an alleviation of the CpG to TpG or CpA mutation.

SNPs (single nucleotide polymorphisms) are thought to show the consequences of base substitution events. These base substitutions are seen to be increased in the promoter region towards the TSS (Guo et al, 2005). Both transversions and transitions increase which is surprising since this region is regarded as highly conserved. The results of CpG transitions are not specifically given though for these SNP data. The odds ratio results in contrast to this suggest that there is a decrease in this type of mutation towards the TSS with respect to CpG and its methylation/deamination products. This is also considered the most common and widespread type of transition substitution. Therefore the two types of analysis seem not to be in agreement. For a full discussion on the odds ratio and how this may relate to substitution event in the upstream see appendix A.13).

Although the directionality of the upstream sequence decreases towards the TSS, the sequence nevertheless possesses a directional bias throughout indicating a general non-randomness in this regard. There are varying structural directional tendencies within the entire upstream sequence studied. This applies for all the dinucleotide pairs. Towards the TSS there is

a change in this structural bias as can be seen as the skew in dinucleotide odds ratio between XpY and YpX, for some dinucleotides.

Strand asymmetry was seen to increase with respect to some of the dinucleotides towards the TSS. Increased asymmetry of the strands towards the TSS with respect to dinucleotide representation cannot be attributed to biased tendencies of nucleotide assembly which form nearest neighbour biases. This is because in theory the tendency to form certain dinucleotides would be identical on both strands. Instead other factors may be used to explain the asymmetry.

In the upstream region transcription coupled repair is not involved since the sequence is not transcribed (Teng et al, 2000). In this region nucleotide excision repair is thought to occur via a different mechanism. This is thought to be different to global repair since in yeast the global repair protein is not required in the upstream. The authors speculate that there is a repair pathway that is a transition between global repair and transcription coupled repair. This transitory pathway or mechanism of repair may contribute to the strand asymmetry towards the TSS observed in this experiment.

The asymmetry of the two strands close to the TSS may be related to transcription and to the regulatory motifs in the DNA sequence. This may result in an increased level of strand recognition by regulatory proteins for transcription so that transcription would occur in the correction rotational orientation. At the promoter rotational orientation is important for the direction of transcription. In contrast, the enhancer may be able to function in either orientation. Furthermore, certain protein binding motifs on the DNA sequence may possibly be the cause of this asymmetry, although this is unknown at present. Therefore it is possible to conclude that the DNA sequence close to the TSS may require strand asymmetry as a means of ensuring that transcription occurs in the correct rotational orientation.

Regarding the bias in structural tendencies between the strands the story is quite different. The enhancement or suppression of dinucleotides as can be seen via the odds ratio skew since this shows that for all the dinucleotide pairs (with only one exception) asymmetry exists in the upstream sequence. However, the level of this skew does not change across the upstream sequence. This means that whilst there are different structural tendencies in the two strands, these tendencies remain similar throughout the region.

2.4.3 Any evidence of boundaries?

This analysis has shown that the sequence compositional features of the 10Kb upstream sequence remain the same for most of this region with the exception of the 4Kb sequence just upstream of the TSS, over which significant changes in dinucleotide composition are observed. Across this 4Kb sequence, varying trends are seen that likely reflect structural and/or functional changes. This approximate location may represent a boundary region (or the upper limit) for regulatory sequences, either including or excluding the enhancer elements which, may occur very far upstream.

The fact that dinucleotide sequence composition changes occur mostly within this 4Kb upstream sequence is useful information since it reveals that this region is different to the further upstream, which is more homogenous for the dinucleotides. Therefore this region probably represents a boundary within the average gene. The 4Kb region of change likely includes the promoter which is usually present at up to 2Kb. However, this sequence probably spans beyond the promoter and may include other regulatory sequences.

For researchers looking for regulatory sequences in the upstream of human genes, for example, as the result of microarray experiments that suggest common gene groupings, this information would be useful. The reason being that it provides a reference point for how far upstream regulatory sequence searches may be carried out.

2.4.4 The effect of repeats

The reason that repeats become increasingly less frequent towards the TSS region is unknown. In order to understand this it is necessary to consider two issues. The first is regarding the possible functions of repeat sequences in the genome. The second is how repeats may be generated in the genomic DNA.

Repeats are thought to be essential for genome function and in the human comprise in total more than 50% of the genomic DNA sequence (Shapiro et al, 2005). One possible role involves the formation of nucleoprotein complexes. Repeat sequences may form boundaries for heterochromatin domains and constitute a significant proportion of matrix attachment regions. This is suggestive of a structural role for these regions of the genome. Additionally, repeats may be involved in transcription regulation. One example of this is observed with the transcription of

a human collagen gene which, is enhanced by dinucleotide repeats (Akai et al, 1999). Also, repeats may even become constituents of protein coding regions. Therefore the notion that these sequences are non-functional has been partially refuted by these examples of their specific roles within the genome. However, the situation is more complex due to different repeat-types.

Another interesting feature of the repeat sequences is that some (in the human and also in other eukaryotic genomes) are thought to form non-B DNA structures in vitro (Brahmachari et al, 1995, Tripathi et al, 1991). For example, (TG/CA)n repeats which are widespread and may play a role in nucleosome organization. In fact, (TG/CA)n and (CT/AG)n are capable of adopting a left handed Z-conformation and may be able to form triple-helical structures.

A general decrease in repeat sequence density over the 5kb region towards the TSS may be due to a decreased need for them in this direction. This may be the result of a relative increase in regulatory protein binding motif density over this 5Kb region. In the same instance there may be a decreased presence of repeat-associated matrix attachment regions in the TSS direction thereby resulting in a reduction of repeats.

Within the 1Kb sequence spanning 5-6Kb upstream of the TSS, repeats in total were present at the 50% level, whereas in the 1Kb immediately upstream of the TSS they were present at the 25% level. Therefore whilst the density is greatly lowered repeats still are present and prominent. In this work only the general repeat density was measured but the type of repeat sequence was not accounted for. So it may be that whilst a majority of repeats are reduced towards the TSS, certain types may be increased. These differences in repeat types would be the subject of further investigation and may correlate with the specific roles attributed to different repeat types. This however goes beyond the scope of this project since repeats are not the main subject of interest. They are only of interest insofar as they may collectively alter the observed trends across the upstream.

The change in repeat sequence density over the 5Kb sequence and the unchanging repeat density further upstream implies and further supports the presence of a boundary around this location. In general the promoter is present at up to 2Kb upstream of the TSS. Therefore this boundary at around 5Kb where change begins towards the TSS, likely indicates the presence of regulatory regions beyond the promoter.

There are different types of repeat ranging from the simple tandem repeat to Alu repeats which contain the highest copy number in the human genome. Simple tandem repeats are thought to have arisen from slippage replication. The decrease in repeat sequence density

over the 5Kb may also or alternatively be indicative of a change in the replication/repair mechanism in this region. However, this seems an unlikely or incomplete explanation for the dramatic change in repeat density, since simple tandem repeats constitute only part of the repeat type repertoire.

A third factor that may be considered in accounting for the reduced repeat density is that this may correlate with the GC content of the upstream sequence which, increases over this region in the TSS direction. Repeat sequences also occur at a lower level in coding sequences which are GC -rich.

Repeats were seen to influence the GC content of the sequence in a counterintuitive manner across the upstream. The repeat fraction actually causes the strong (SpS) dinucleotides to increase further upstream and the weak (WpW) dinucleotides to decrease in this direction. In the event that the composition of the repeats would change across the upstream region, the expectation is for the change to occur in sink with the rest of the sequence and not in direct opposition. The reason for this is unknown, but it does imply a different design and role for the repeat regions. This may be a means via which certain repeats are distinguished against the background sequence across the different locations of the upstream.

Repeat sequences (or their elimination) had some effect on dinucleotide composition and representation. They changed the composition of dinucleotides; however relative values remained the same. Most importantly repeats did not alter the general dinucleotide trends across the upstream sequence. This may seem surprising since the repeats constitute a large proportion of the sequence and this proportion is greatly diminished over the 5Kb sequence upstream of the TSS.

Regarding the dinucleotide odds ratio, filtering out repeat sequences made the profile less random, although this was not true for all dinucleotides. This makes sense in light of what is already known about repeats and overall sequence functionality. Repeats are relatively simple sequences with what seems to be a lower level functionality in many cases. Therefore it is expected that their increased presence should confer more random-like characteristics on the sequence. Having said this, relative dinucleotide odds ratio values were not dramatically altered, which highlights the stability of dinucleotide relative abundance profile even with respect to repeats.

Determination of the significance of repeats in the human has become an important focus in genome sequence studies. In this work repeat fractions were not initially eliminated from the upstream sequence. Instead the upstream sequence was considered with and without

repeats. In conclusion, whilst repeat sequences had a slight effect on dinucleotide sequence features they did not alter the overall trends and changes seen across the upstream region of the human gene.

2.4.5 Limitations of the dataset and the experiments

The upstream sequence dataset

In the first place the results obtained in these experiments depend on the quality of the sequence data and database annotation. The human genome has been extensively sequenced and there is a large data set of genes allowing for a large sample size, although the real gene number remains unknown. Gene prediction algorithms have been largely unsuccessful and the problems include; knowing where the gene begins, identifying the TSS and determining exon/intron boundaries.

This situation is further complicated by the existence of alternative splicing and also alternative start sites in mammalian genes (Tran et al, 2002, Yamashita et al, 2006). Therefore when mining the human genome for the 5' upstream, intronic and exonic sequences there is an unknown margin of error. An attempt was made to limit this error-margin though by utilising sequences derived from known or verified gene sequences. Also, it is possible that the 10Kb 5' upstream sequence in some cases may coincide within an unidentified upstream gene.

An important limitation is that it is unknown how far upstream the cDNA's in the genome actually extend. I.e. the exact location of all the 5' termini has not been accurately identified. In this project the 5' upstream sequence was taken to begin at the TSS. Therefore it may be that a proportion of the upstream sequence used in this project contains some 5'UTR or that the TSS was taken to be further upstream than the true start site.

To determine the extent of this problem, an oligo-capping method has been developed (Suzuki et al, 2002) to obtain full-length cDNA's for the human. The resulting sequences were then compared with entries in the RefSeq database. The authors estimate that 34% of sequences should be extended towards the 5'end. On average the extension was 83 bases. This describes the possible margin of error.

Although this is a limitation, it is unlikely to dramatically shift dinucleotide changes observed in this work. Even if the 250 base sequence closest to the TSS would be excluded from the analysis, the compositional changes observed would still apply, since they were observed over a far larger region. This uncertainty regarding the TSS, though does introduce inaccuracy.

The dinucleotide representation (odds ratios)

Although the 1Kb (and 250 base) sliding windows may not be accurately fixed across the dataset of genes and there may be frame shifts in position due to reasons of inaccuracy that have already been discussed. Therefore the positional location with respect to the TSS is not entirely accurate and homogenous across the dataset. Also, there may be considerable variation across the different genes with respect to positional aspects of sequence, for example the promoter is not found in a fixed place in all genes. However, in all likelihood the promoter will occur within a certain range (of positional location) within the upstream.

Also, the length of the sequence segments of 1Kb (and 250 base) was of course arbitrary, and there would be many different ways to design such an experiment. Very specific regions (such as the TATA box and flanking sequences etc...) within the upstream have not been considered rather the results are only a view of general changes over the 2Kb and 10Kb regions of the upstream. Therefore changes in local sequence and structure are unknown and not taken into account here. It is important to remember that the results for changing trends in sequence composition across the upstream are average results within each dataset and are only intended to represent an 'average' situation.

Dinucleotides and structure

It has been assumed that an increase in a particular dinucleotide correlates with an increase in the associated structural features of the DNA. Whilst structures of DNA possessing certain dinucleotides have been ascertained via X-crystallography, the overall structure of any given region of DNA depends on specific sequence rather than the dinucleotide composition. Therefore the structural trends described here are only general changes and are not relating to specific DNA structure, which is far more complex. Furthermore the odds ratio or representation of a dinucleotide and any structural tendency described relates only to a 'tendency' and not to an actual structure.

Repeat masking

The masking used here did not necessarily account for all human repeats. The subject of repeat sequences in human DNA is complicated and there may be unknown repeat sequences that were unaccounted for in this analysis. The conclusions drawn must therefore bear this in mind.

Why were repeats not filtered out in the first place? The repeat sequences are an integral part of the DNA. Within the upstream they may play a structural role and may be involved in the process of regulation (Iglesias et al, 2004, Iyer et al, 1995). The problem with removing them is that a potentially important component of the upstream sequence is removed. Although often the assumption is that repeats are superfluous material and they are often filtered out in such experiments, this may not be the case. Removing then may lead to a loss of valuable information. For example, a significant over-representation of transcription factor binding sites has been found in both repeat-containing and repeat-filtered sequences of mammalian genomes (Stepanova et al, 2005). In the same instance, it is important to know the effect they have on the upstream sequence and in the context of these experiments how they may affect changes in sequence trends across the upstream.

2.4.6 The overall message and questions that arise

In summary this work so far, has identified changes in dinucleotide sequence composition and representation across the upstream sequence of the human gene which, reflect structural and functional features of this region. An approximate boundary region for which dinucleotide changes were no longer seen was located. The boundary idea was further supported by a similar observation for repeat density changes.

Structural changes were observed including decreased flexibility and suppression of flexible dinucleotides towards the TSS. Also, bistable dinucleotide composition was greatly increased towards the TSS. This dual alteration in dinucleotide composition and representation is likely important for transcription regulation.

Sequence directionality exists across the entire 10Kb upstream sequence with respect to dinucleotides. This is probably very much related to the directional nature of the sequence

because of nearest neighbour nucleotide tendencies and mutational tendencies. More specifically in the upstream sequence the directional nature may be related to transcription and protein interaction. Counter to this idea, the directionality is decreased towards the TSS, possibly due to an alleviation of transition mutations. However, there is a discrepancy between this result and SNP data. This issue remains unresolved.

Dinucleotide compositional asymmetry generally increases across the 10Kb upstream towards the TSS. The result of this would be increased strand recognition by regulatory proteins for transcription that would be a possible determinant of the correct direction for transcription.

Finally, repeat sequences were found not to alter the changes or trends in dinucleotide composition and representation across the upstream. Their density though is greatly decreased towards the TSS and their composition is different to their surrounding sequence. This implies location specific roles for repeats within the upstream and some level of independence of composition and structure.

In eukaryotes DNA sequences in close proximity on the molecule can have very different roles. Therefore the changes seen in and across the 10Kb 5' upstream sequence of the gene are to be expected. Changes or fluctuations in local sequence and structure are not considered here. Rather, this method served the purpose of investigating general dinucleotide trends along the region. Also, the gene sequence datasets were not divided into those containing different promoter types, etc... This may be subject of further work.

Several paths may be followed in order to continue this work. Examples include a more detailed analysis of dinucleotides and larger motif similarities to clarify the 4-5Kb upstream boundary region. Regarding repeat sequences there is much scope for investigation. An analysis including a division into different repeat-types and their changes in density and composition across the upstream may be carried out as a means to understanding their role in this region of the genome.

Furthermore separate datasets of genes may be taken with different promoter types in order to make boundary comparisons. Differences in dinucleotide composition and representation between regulatory and spacer sequence within the promoter and also enhancer elements may be studied.

This chapter has shown some simple trends or changes across the upstream sequence of the human gene. Although several different areas of further work have been outlined above, the next step of investigation involves a specific area which has been highlighted here. The

upstream sequence was shown to possess non-random characteristics with respect to dinucleotides. These trends were then be subdivided according to the purine-pyrimidine or weak-strong content of the dinucleotides. In the experiments that follow the overall R/Y and W/S non-random features of the sequence and any changes that occur across the upstream in this regard will be studied.

3. Distance from Randomness of the Upstream Region of the Human Gene: The Real Sequence versus the Random Model

3.1 Introduction

In the previous chapter dinucleotide changes were observed across the 5' upstream sequence of the human gene. The composition and representation of individual dinucleotides was analysed. This revealed information about structure, directionality and strand asymmetry. In this chapter the overall distance from randomness of the upstream (and other genomic regions) is analysed. In other words, the real sequence is compared in its entirety to a randomised sequence model. This measure shows in a general sense the non-random nature of the genomic DNA and how this varies along sequences in close proximity.

One of the observations was that weak and strong properties change across the upstream. This was in particular regarding dinucleotide composition. Also, purine and pyrimidine features changed for dinucleotide representation. In this chapter, these specific features of the sequence are the subject of further investigation.

3.1.1 Random and non-random sequences: levels of functionality

It has already been discussed that randomised sequences contain no common specific structure or function. Therefore if DNA sequence properties were analysed, they would be expected to possess different properties to randomised sequences of equivalent composition (Blaisdell, 1983).

The set of dinucleotide representation profiles may be described as a genomic signature since this profile describes how different the sequence is to randomness and may be used to compare different sequence types, each one possessing its own unique signature.

The under- or over- representation of dinucleotides and the resultant genomic signature that takes into account the set of all possible dinucleotides is determined via two different mechanisms (Karlin et al, 1998). In general significant under- or over-representation of motifs

in the DNA are indicative of functional elements. These may be the result of either context dependent mutation or a selection for structural features of the DNA.

Meaning of the sequence being relatively distant or close to randomness

A change in a dinucleotide from a random score to a less random score would imply an altered use for that dinucleotide across the upstream. If a dinucleotide is relatively more distant from the random model intuitively its significance seems greater in the sequence. This significance may be due to either its relative enhancement or suppression, i.e. its relative absence or presence is of importance.

In theory, if the sequence is very close to randomness with respect to all the sixteen dinucleotides, it may be regarded as a close to random sequence. This implies a relatively lower level of functionality. Relative changes in distance from randomness across the 10Kb upstream sequence (and other genomic regions) would provide information about differences in sequence 'design'.

The expected results are that all real DNA sequences would be non-random. When cross-comparing different types of sequence, the expectation is for non-coding sequences to be closer to the random expectation than coding sequences since coding sequences are considered highly functional as they code for proteins and contain the information for protein motifs. Across the 10Kb upstream the sequences closer to the TSS (since they contain the promoter and regulatory elements) are expected to be more distant from randomness than the further upstream sequence.

3.1.2 Division into two categories; Purines and pyrimidines versus weak and strong hydrogen bonding bases

Subdivisions into different base properties

The four different bases of the DNA molecule each possess their own properties. Adenine contains a two-ring structure and has the potential to form two hydrogen bonds. Thymine contains one-ring structure and also has the potential to form two hydrogen

bonds in the base pair. Cytosine has one-ring and the potential for three hydrogen bonds. Guanine has two-rings and forms three hydrogen bonds. Therefore within the bases there are two different categories of property that define the difference or variation between them.

This apparently simple variation is of great importance for the role of DNA in the cell. For example, the genetic code relies on the inherent differences between the bases (which the cell machinery is able to recognise) in order to form that chemical code. Also when proteins bind to the DNA (as occurs for transcription) the protein must be able to recognise the difference and distinguish between the bases in order to make the binding specific and meaningful. Therefore these two properties can be regarded as generating the 'combination' of the four bases, each one being unique and recognisable by the cell. It is obvious that this ring structure and hydrogen-bonding also explains the logic behind the standard base pairing.

The bases may be described within two separate categories; as weak (A and T) and strong (C and G) bases. A 'weak' hydrogen bonding base is capable of forming two hydrogen bonds in the base-pair whereas a 'strong' base forms three hydrogen bonds. Alternatively, they may be described as purines (A and G) or pyrimidines (C and T). These are two different ways to group the base, which means that in theory the sequence may be viewed as comprising either of; 1. purine (R) and pyrimidine (Y) bases or alternatively as 2. weak (W) and strong (S) bases. The properties of these two categories will be discussed.

The effect of hydrogen bonding on melting temperature of DNA

The weak/strong base sequence of the DNA determines its melting point which, may change at different locations on the molecule. This is due to the two hydrogen bonds that exist between the A-T base pair and the three hydrogen bonds between the C-G base pair. The higher the content of strong bases the higher the melting temperature. AT rich regions constitute the classical DNA unwinding motif. Over-representation of weak (AT-rich) long tracts has been found (Shomer et al, 1999) in bacteria as well as their tendency to unwind, thereby suggesting a role in promoter function.

Hydrogen bonding patterns between amino acid and DNA base-pairs for protein-DNA binding

Potential hydrogen bonding patterns within the base-pairs of the DNA is essential for direct readout by regulatory proteins. The base pairs possess different patterns of hydrogen bond acceptors and donors that potentially form bonds with the amino acid side chains (Seeman et al, 1976) (see figure 3.1). These patterns may constitute a recognition code that is dependent

upon the DNA sequence. Furthermore the major and minor grooves of the helix are different in their potential to form hydrogen bonds with amino acids.

The way in which the amino acids probe the specific base-pair sequence is unknown. Although the location of H-bond donors and acceptors are known in the base-pairs, the recognition pattern for this direct readout by the amino acids remains undetermined. However, it is still possible to discuss the potential patterns of H-bond donors and acceptors within the base-pairs (Hoglund et al, 2004).

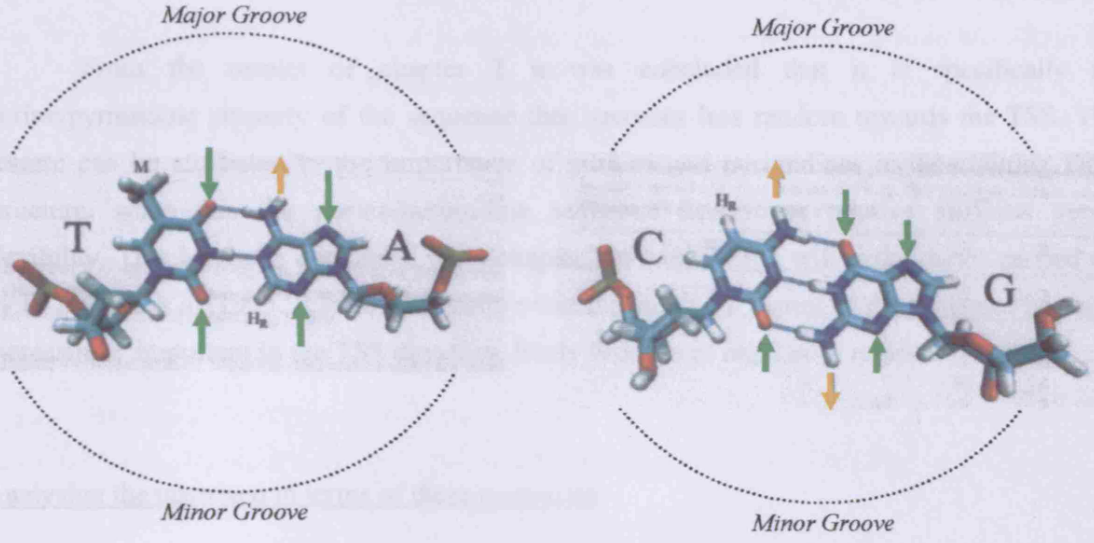


Figure 3.1: Diagram of recognition patterns for hydrogen bond donors and acceptors in the base-pairs (based on diagram by Hoglund et al, 2004).
These patterns are shown for the major and minor grooves of B-DNA.
H-bond acceptors are shown by green arrows and donors are shown by yellow arrows.
In the minor groove it is only possible to distinguish a C-G base-pair from a T-A base pair via these H-bonding patterns. In the major groove it appears possible to distinguish all four bases.

The major groove G-C base-pair possesses the following pattern for potential H-bond formation from the G; a hydrogen bond acceptor followed by a second acceptor, and hydrogen bond donor. When the base pair begins with a C, this pattern is reversed in relative orientation. The A-T base-pair has the following recognition code starting from A; a hydrogen bond acceptor, a donor, and another acceptor. Here though when the base-pair begins with T the order of the pattern is the same as when it begins with an A. Within a sequence these can create a pattern for recognition by the protein via H-bonding.

The minor groove possesses a simpler recognition code. For G-C base pairs the pattern is as follows; a hydrogen bond acceptor, a hydrogen bond donor and another hydrogen bond acceptor. For A-T base-pairs there are two hydrogen bond acceptors. These minor groove patterns appear identical in either orientation. This means that within the minor groove of the DNA sequence the weak/strong sequence is the determinant of potential hydrogen bonding patterns, whereas the purine/pyrimidine sequence does not determine this pattern.

The purine/pyrimidine sequence is a greater determinant of DNA structure

From the results of chapter 2 it was concluded that it is specifically the purine/pyrimidine property of the sequence that becomes less random towards the TSS. This feature can be attributed to the importance of purines and pyrimidines in determining DNA structure, since this the purine/pyrimidine sequence determines relative stiffness versus flexibility. This has been concluded from chapter2 in conjunction will experiments carried out by El Hassan et al, 1995. It follows therefore that the structural aspect of the sequence becomes increasingly important in the TSS direction, likely because of regulatory regions.

Analysing the upstream in terms of these properties

An analysis of DNA in terms of the two different sub-divisions of base property, namely; purines/pyrimidines (R/Y) and weak/strong (W/S) bases may provide a better understanding of the role of the four bases in accordance with these divisions of their properties. The question is: how important across different locations of the 10Kb upstream region is the hydrogen bonding capacity of the bases, which is characteristic of the weak/strong sequence? Alternately, how important is size and ring structure that is characteristic of purines/pyrimidines and their resultant effects on DNA structure? These questions can only be addressed in a relative sense, i.e. one upstream location (close to the TSS) compared with another or the upstream sequence compared with the coding sequence etc...

Another way to look at this is to pose the following hypothetical question; to what extent is the choice of base in the DNA sequence dependent on whether it is a purine or pyrimidine, versus a weak or a strong base? It must be that both the structure and potential to form hydrogen bonds is important for their role the DNA sequence. To what extent is each of these important, and in which context? In the upstream, it is of interest to find out to what extent the choice of base is dependent upon these factors and how these may change across the different positional upstream locations.

3.1.3 Aims and experimental design

The upstream sequence and its purine/pyrimidine and weak/strong translations

This analysis of purines/pyrimidines and weak/strong bases was carried out due to the observation (in the previous experiments, chapter2) of subdivisions in the frequency and representation of dinucleotides into these two separate categories. The next step was to work out distance from randomness utilising the genomic signature (described below) in a sequence that was regarded only as a purine and pyrimidine entity. This was then repeated for weak and strong bases and a cross comparison made. Changes in sequence properties analysed across segments of the 10Kb upstream sequence of the human gene as was done in chapter2.

In order to analyse these (purine/pyrimidine versus weak/strong) nucleotide properties separately, the ATCG upstream sequence was translated into two different but equivalent sequences. The first was a purine/pyrimidine (R/Y) sequence and the second was a weak/strong (W/S) sequence. This resulted in two separate translated sequences which were treated as separate entities for analyses (see figure 3.2). This was done so that the relative importance of these two subdivisions of nucleotide properties could be assessed individually.

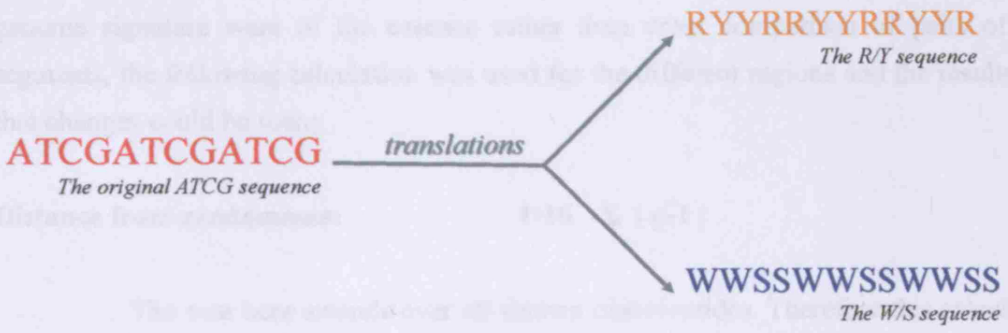


Figure 3.2: Diagram depicting that in all analysis of this chapter the original (real) ATCG sequence was translated into two equivalent sequences.
The first sequence translation was for purines and pyrimidines, so that; (i) A and G were converted to R and (ii) C and T were converted to Y. This sequence was then referred to as the R/Y sequence.

The second sequence translation was for weak and bases, so that; (i) A and T were converted to W and (ii) C and G were converted to S. This sequence was referred to as the W/S sequence.

These translations permit a single sequence to be analysed for two separate categories of base property.

The genomic signature

A measure of distance between two sequences either within or across organisms has been referred to as the average absolute dinucleotide relative abundance difference (Karlin et al, 1996, Karlin et al, 1997). This has been used to cross compare different genomic signatures and is calculated is as follows:

Average absolute dinucleotide relative abundance difference:

$$\delta(f,g) = 1/16 \quad \Sigma \; | \rho(f) - \rho(g) |$$

In this equation, **f** is one sequence type whereas **g** is another. For example, **f** may be a human sequence whereas g a mouse sequence. **ρ(f)** is the odds ratio value for a dinucleotide for one sequence type. **ρ(g)** is the odds ratio for the same dinucleotide within the second sequence type. The sum here extends over all sixteen possible dinucleotides. See section 2.1.6 for details of the odds ratio.

In this project, sequence changes across the ten different positional datasets (*upstream1*-to-*upstream10*) of upstream region were of interest. Since changing trends in the genome signature were of the essence rather than cross comparison of pairs of individual segments, the following calculation was used for the different regions and the results plotted so that changes could be seen;

Distance from randomness: $1/16 \quad \Sigma \; | \rho - 1 |$

The sum here extends over all sixteen dinucleotides. Therefore this calculation is the average of odds ratio values for all sixteen possible dinucleotides for a particular sequence. Also, this calculation does in fact compare two types of sequence. However, instead of comparing two real genomic DNA sequences, a real sequence is compared with its random equivalent, since the theoretical odds ratio value for the random sequence is 1.0 (Karlin et al, 1998). Therefore this value may also be regarded as an average deviation of dinucleotides from the expectation for a random sequence of equivalent mononucleotide proportions.

This calculation has been referred to as the 'total distance from randomness' of the sequence in this project. The further this value is from zero, the further away the value is from the random (or shuffled) sequence. This permitted each of the upstream datasets (*upstream1*-to-*upstream10*) to be seen relative to its random expectation. This total distance from randomness value was calculated for each of the ten upstream datasets of sequence, for *intron1*, *exon1*, *coding1* and also for the *whole genome*.

Adaptation of genomic signature of R/Y and W/S sequence translations

The dinucleotide proportions, odds ratios and genomic signature calculations were then carried out for R/Y and W/S upstream sequences datasets as described above for the original (ATCG) datasets. This time though the dinucleotides were changed accordingly. Therefore, for the purine/pyrimidine datasets, the dinucleotides would be; RpR, RpY, YpR and YpY, i.e. four possible dinucleotides instead of sixteen. For the weak/string dataset the dinucleotides were; WpW, WpS, SpW and SpS. Therefore the distance from randomness calculation was as follows;

Distance from randomness: $1/4 \quad \Sigma \mid \rho - 1 \mid$

The sum here extends over all four possible dinucleotides.

3.2 Methods

The DNA sequences for this project were obtained from the NCBI human genome database, build 35. The 10kb 5' upstream sequence of the gene was utilised. Also, sequences from the first exon, the first intron and also the genome-wide sequence were used in the analysis. These human genomic DNA sequences were identical to those used in chapter 2 (see methods section 2.2.1 for details). As in chapter 2, The 10Kb upstream sequence sub-divided into ten 1Kb non-overlapping portions; *upstream10-to- upstream1*. The dataset named *upstream1* being just adjacent to the TSS.

3.2.1 The genomic signature: distance from randomness

Within each upstream dataset, e.g., *upstream1*, distance from randomness values (**1/16 Σ | ρ-1 |)** were calculated for each individual 1Kb sequence fragment taking into consideration its specific nucleotide and dinucleotide composition. Distance from randomness results were then averaged (using the median) over the entire dataset of 18,725 sequence fragments. The same was carried out for each of the ten upstream datasets; *upstream1-to-upstream10*, *intron1*, *exon1* and *coding1* and also for the *whole genome*.

For the genome-wide sequence, the distance from randomness values were taken across all the contigs in of each individual chromosome. E.g. the frequency of dinucleotide XpY and nucleotides X and Y were found in the entire sequence of chromosome 1. The odds ratio could then be worked out and then the distance from randomness calculation could be made. A mean value was calculated for distance from randomness across all the chromosomes.

3.2.2 The R/Y sequence translation versus the W/S translation

The genomic DNA sequences include all upstream datasets; *upstream10-to- upstream1*, first exon, the first intron and also the genome-wide sequence were subjected to two separate translations. The first is the translation of the original ATCG sequence to a purine/pyrimidine (R/Y) sequence and the second is the translation of the original sequence into a weak/strong (W/S) sequence. This yielded two sequence datasets in addition to the original ATCG dataset

for the genomic DNA sequences. Distance from randomness values (**1/4 Σ | ρ-1 |**) were worked out for these two (R/Y and W/S) sequence translations in the same way as described above for the original sequences.

3.3 Results

3.3.1 Relative distance from randomness across the upstream sequence

The results show that the 5' upstream sequence becomes closer to randomness towards the TSS (see figure 3.3). The distance from randomness score remains fairly constant for 7Kb of the sequence; *upstream10*-to-*upstream3* distance from randomness is between 0.189-0.190. This value then decreases to 0.169 in *upstream1*. Therefore this change occurs across the 3Kb sequence closest to the start site. This result is unexpected since this region has a high density of regulatory elements and may therefore be considered a more highly functional region.

The result for the other genomic regions was also unexpected, with the intronic sequence (*intron1* distance from randomness; 0.193) being the most distant from randomness. Also, the difference between the *upstream1* and *upstream10* (therefore the change across the 10Kb upstream) is greater than the difference between the *upstream1* and *coding1* with respect to these values (see figure 3.4).

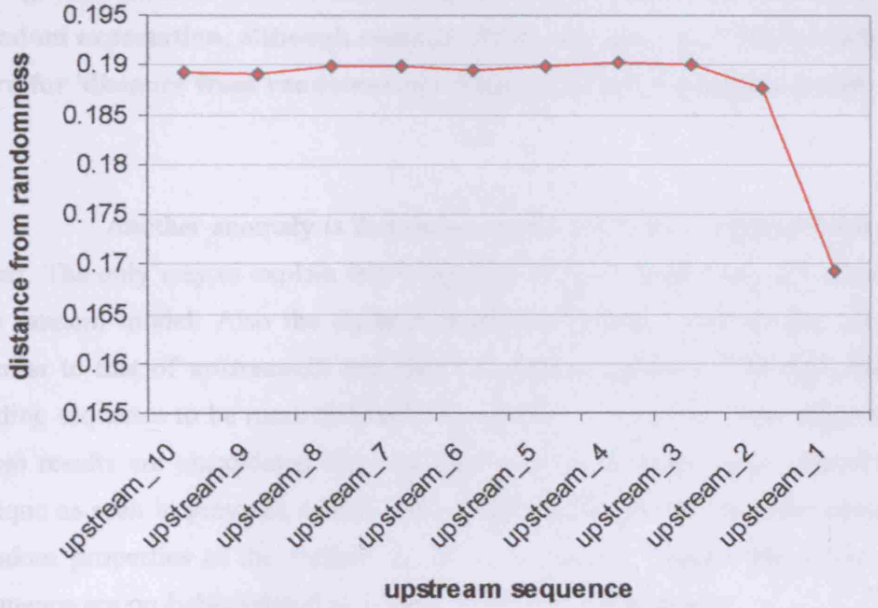


Figure 3.3: Distance from randomness (1/16 Σ | ρ-1 |) graph for the 10Kb upstream sequence.

This graph shows that the 5' upstream sequence becomes closer to randomness towards the TSS. The distance from randomness value remains fairly constant for 7Kb of the sequence (*upstream10*-to-*upstream3*) and then decreases for the x3 (1Kb) sliding-windows

closest TSS (*upstream3*-to-*upstream1*). A value of zero for 'distance from randomness' is equivalent to the random model.

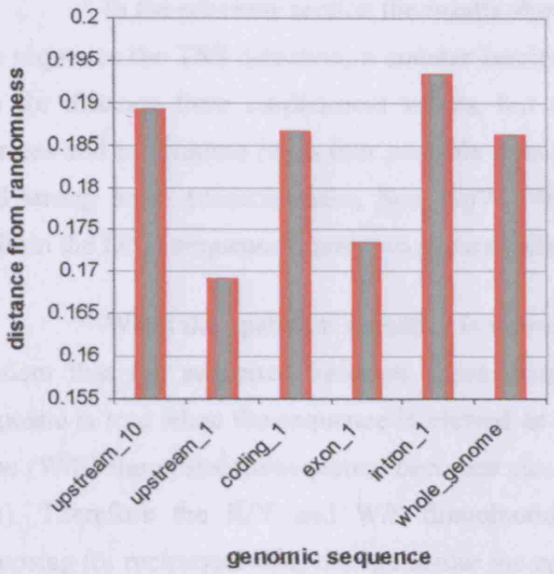


Figure 3.4: Distance from randomness (1/16 Σ | ρ-1 |) graph for different genomic regions.
The result for the other genomic regions was also unexpected, with the intronic sequence being most distant from randomness. The *exon1* region also is surprisingly close to the random expectation, although *coding1* which excludes the UTR is much higher. A value of zero for 'distance from randomness' is equivalent to the random model.

Another anomaly is that these exonic and coding sequences are so different to each other. The only way to explain this is that the UTR has made the exon sequence much closer to the random model. Also the distance from randomness value for the *whole genome* is very similar to that of *upstream10* and also the coding sequence. The expected result was for the coding sequence to be more distant from randomness than the non-coding sequence. Although these results are unexpected they are interesting in that the region closest to the TSS appears unique as seen in previous results. This is true insofar as the sequence possesses different non-random properties to the further upstream sequence. Changes that occur across a stretch of sequence are probably related to varying structure and function.

3.3.2 Distance from Randomness: Sequences viewed as purines/pyrimidines versus weak/strong

In the previous section the results showed a decrease in distance from randomness of the sequence the TSS direction; a counter-intuitive result. In this section the results presented are for distance from randomness values, but this time viewing the sequences as either; 1. Purines and pyrimidine (with four possible dinucleotides, RpR, RpY, YpR and YpY), 2. Weak and strong bases (dinucleotides, SpS, SpW, WpS and WpW). Hence this divides the bases (within the DNA sequences) into two separate classes.

When the upstream sequence is viewed as consisting of purines and pyrimidines it is evident that the sequence becomes more distant from randomness towards the TSS. The opposite is true when the sequence is viewed as consisting of weak and strong residues. In this case (W/S) the upstream sequence becomes closer to randomness towards the TSS (see figure 3.5). Therefore the R/Y and W/S dinucleotide distance from randomness values display opposing (or reciprocal-like) change across the upstream sequence.

In general, when the different genomic sequences are viewed as consisting of purines and pyrimidines, the non-coding DNA is more distant from randomness than coding DNA (see figure 3.6). Distance from randomness values for the non-coding are as follows; *upstream10*=0.122, *upstream1*=0.134, and *intron1*=0.148. The entire genomic DNA which is mostly non-coding has a value of 0.118. The coding DNA have the following distance from randomness values; *coding1*=0.103 and *exon1*=0.118.

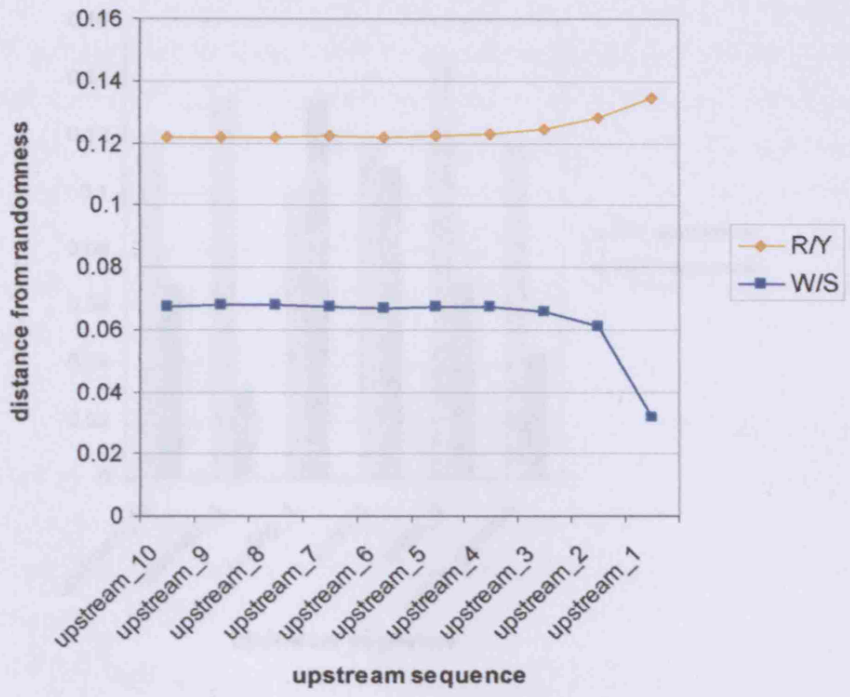


Figure 3.5: Distance from randomness (R/Y and W/S) in the upstream sequence.
The results are shown for the two subdivisions of the four nucleotides into i). Purines and pyrimidines, ii) Weak and strong bases.
The results show that the distance from randomness values remain about constant from the *upstream10*-to-*upstream5* segment datasets for both R/Y and W/S. Between *upstream5* and *upstream1* there is an increase in distance from randomness with respect to R/Y, and a decrease with respect to W/S.
Therefore the R/Y and W/S arrangement within the sequence show opposite distance from randomness trends towards the start site region.

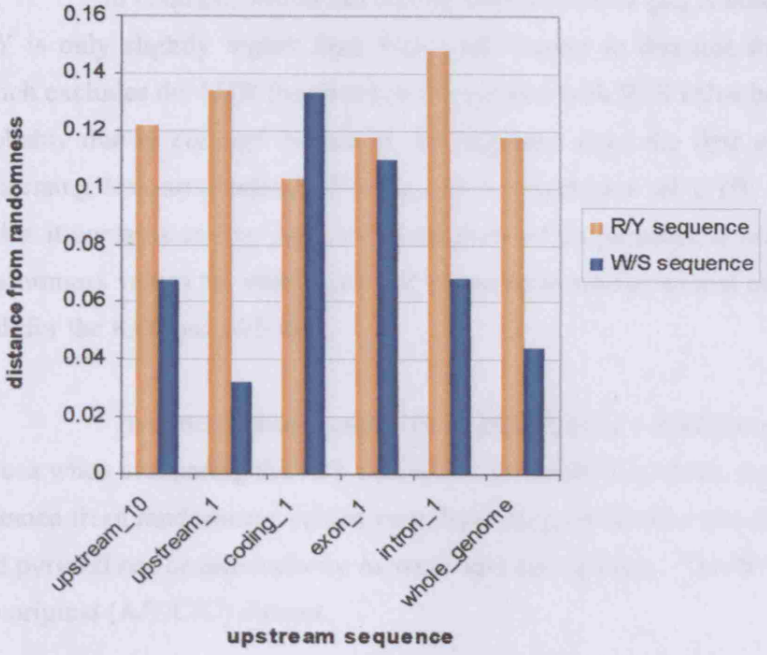


Figure 3.6: Distance from randomness for the different genomic (R/Y and W/S) sequences. The profile shows the average distance from randomness of the different genomic sequences, including; *upstream10*, *upstream1*, *coding1* (the first exon excluding the UTR) *exon1* and the *whole genome*.

From the R/Y perspective non-coding DNA is more distant from randomness than coding DNA. For W/S the coding DNA is more distant from randomness than the non-coding DNA. In this sense R/Y and W/S display opposing qualities.

Within non-coding DNA, R/Y sequences are much more distant from randomness than the W/S sequences.

This shows that there is a division of distance from randomness when comparing the R/Y and W/S equivalent sequences, depending on whether the sequences are coding or non-coding.

When the sequences are viewed as weak and strong bases the opposite is true; the coding DNA is more distant from randomness than the non-coding DNA. Distance from randomness values for the non-coding are as follows; *upstream10*=0.067, *upstream1*=0.032, and *intron1*=0.068. The *whole genome* =0.044. The coding DNA have the following distance from randomness values; *coding1*=0.133 and *exon1*=0.110. In this sense R/Y and W/S possess opposite trends regarding coding and non-coding DNA sequences. Also, within non-coding DNA, the R/Y sequence is much more distant from randomness than the W/S sequence.

In contrast, within the coding sequences this gap is comparatively smaller. In *exon1* R/Y is only slightly higher than W/S with respect to distance from randomness. In *coding1* which excludes the UTR the situation is reversed with W/S value being higher than R/Y. This is probably due to *coding1* (which is the sequence from the first exon that excludes the UTR) possessing less non-coding DNA due to the exclusion of UTR. The genome-wide sequence whilst it contains coding regions the majority of its sequence is non-coding. The distance from randomness values for entire genome sequence is similar to that of the other non-coding DNA, both for the R/Y and W/S data.

In general these results show that there is a division of distance from randomness values when comparing the R/Y and W/S equivalent sequences. I.e. in the same genomic DNA, distance from randomness values vary depending on whether the sequence is viewed as purines and pyrimidines or alternatively as weak and strong bases. The W/S profile is similar to that of the original (A/T/C/G) dataset.

All of the distance from randomness trends viewed so far across the upstream region are for sequences that are unmasked for repeats. In the repeat-free sequence trends of dinucleotide distance from randomness are the same as for repeat-containing sequence. This is true for the R/Y, W/S and original (ATCG) sequence. The essential difference between repeat masked and unmasked sequences is that repeats make the sequence more random. For a full comparison of masked and unmasked data see appendix B.1 and B.2. These results are not presented here since there is no difference in the overall trends across the 10kb upstream sequence.

3.3.3 Representation of the individual R/Y dinucleotides

As already seen the R/Y upstream sequence becomes more distant from randomness towards the start site of transcription. Also, each of the four possible dinucleotides individually becomes more distant from randomness (see figure 3.7). RpR and YpY are over-represented throughout the 10Kb upstream, and they become more over-represented towards the TSS. In contrast, YpR and RpY are both under-represented throughout the 10Kb sequence, and they become more under-represented.

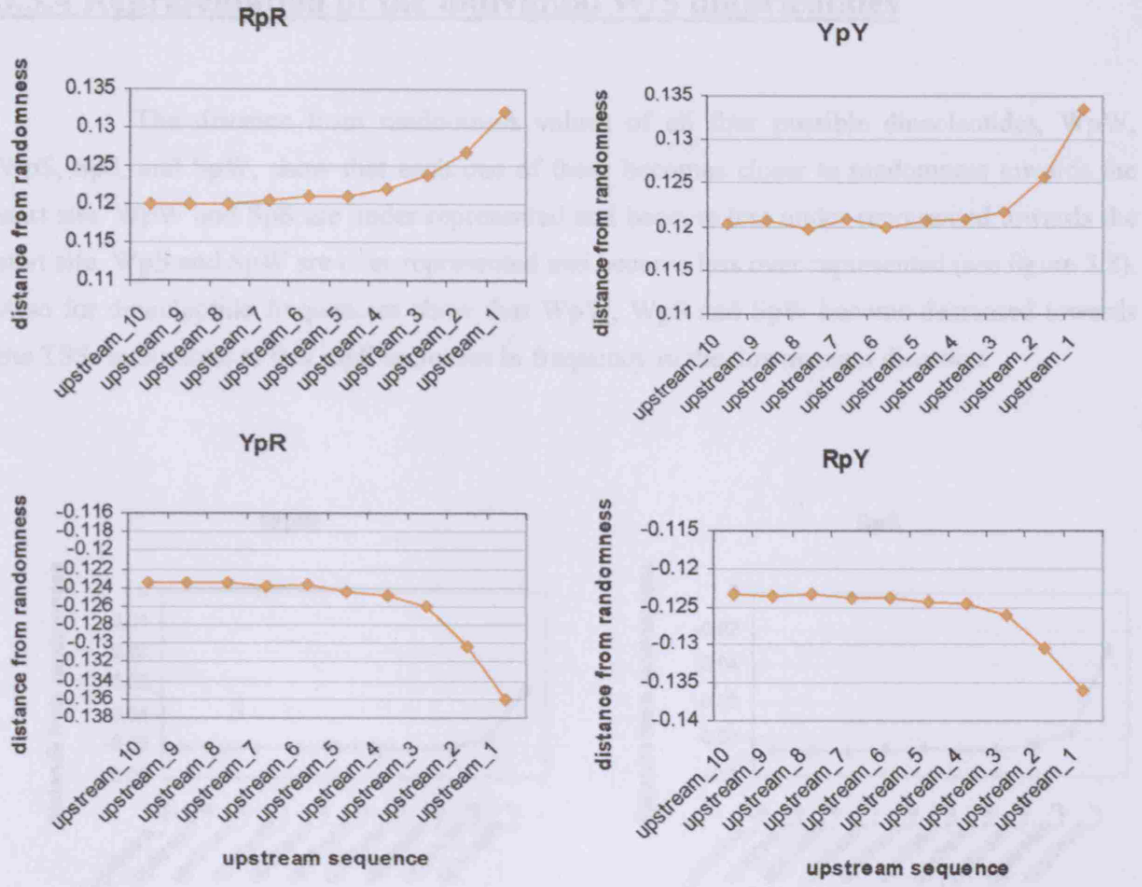


Figure 3.7: Distance from randomness charts for individual R/Y dinucleotides across the upstream

These graphs and data-table are a breakdown of the distance from randomness value across the 10Kb upstream. Here the representation values are given for each of the four possible dinucleotides; RpR, RpY, YpR and YpY. The value for a random sequence would be zero, to which the odds ratio values may be compared. The results show that the sequence becomes less random for each of the four dinucleotides towards the TSS.

The dinucleotides RpR and YpY are over-represented throughout the upstream sequence and become more over-represented toward the TSS. RpY and YpR are under-represented throughout the upstream and become more under-represented in the downstream direction.

3.3.4 Representation of the individual W/S dinucleotides

The distance from randomness values of all four possible dinucleotides, WpW, WpS, SpS, and SpW, show that each one of these becomes closer to randomness towards the start site. WpW and SpS are under-represented and become less under-represented towards the start site. WpS and SpW are over-represented and become less over-represented (see figure 3.8). Also for dinucleotide frequencies show that WpW, WpS and SpW become decreased towards the TSS. In contrast to this, SpS increases in frequency in the downstream direction.

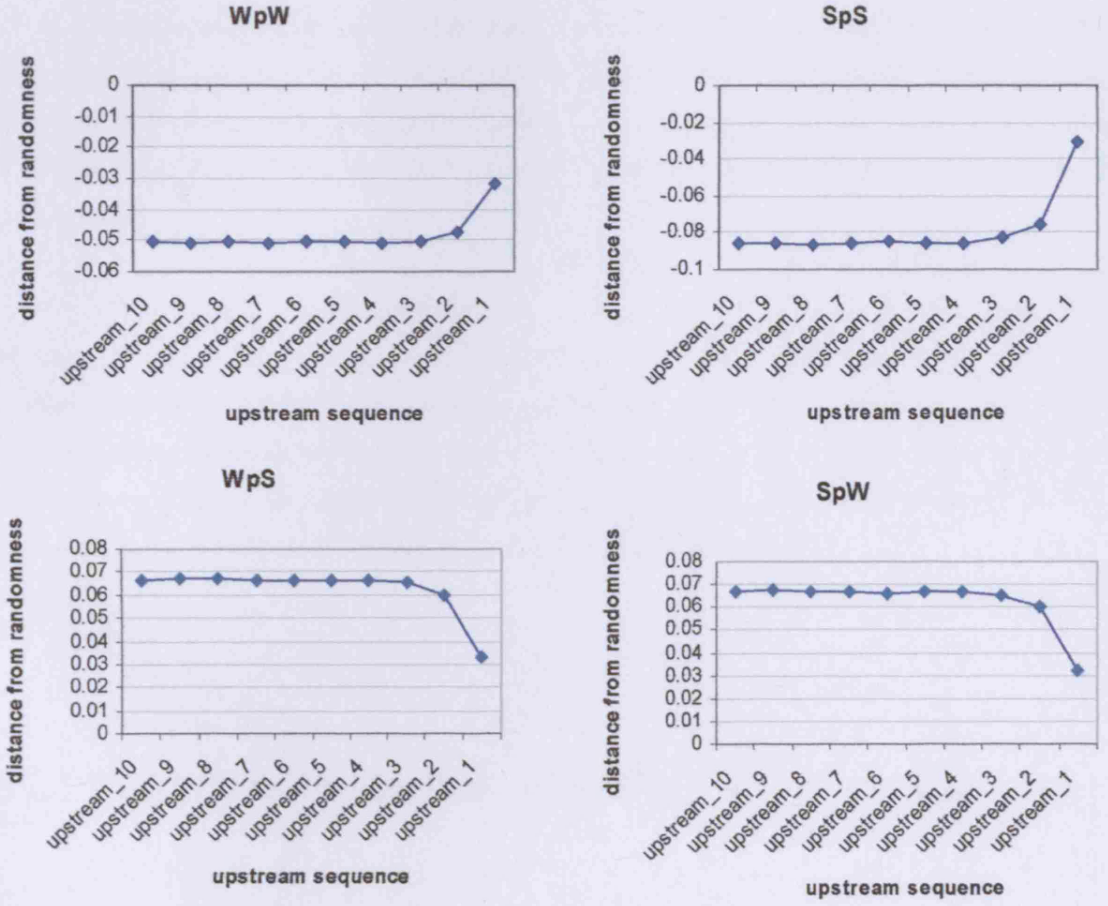


Figure 3.8: Distance from randomness charts for individual W/S dinucleotides across the upstream

These graphs and data-table are a breakdown of the distance from randomness value across the 10Kb upstream W/S sequence. The results show that the sequence becomes closer to the random model for each of the four dinucleotides towards the TSS.

The dinucleotides WpS and SpW are over-represented throughout the 10Kb upstream sequence and become less over-represented toward the start site. SpS and WpW are under-represented throughout the upstream and become less under-represented.

Repeat free upstream sequences display similar trends for R/Y and W/S dinucleotide composition and distance from randomness as those seen for the unmasked sequences. For a comparison of equivalent repeat masked and unmasked dinucleotide composition across the upstream sequence see appendix B.3 and appendix B.4. For a comparison of dinucleotide representation see appendix B.5 and appendix B.6.

3.4 Conclusions & Discussion

3.4.1 Meaning of differences across the ATCG sequence

What does the change in distance from randomness actually mean for the upstream sequence, its structure and function? In theory a set of sequences that have a particular function should be more distant from randomness than sequences that do not possess function, due to a need for specific motifs.

Presumably in a highly functional sequence such as the coding region which possesses its own language, the arrangement of bases would be biased towards a sequence that makes sense. This bias would make it less random than a sequence that possesses no language at all. So the more highly functional the sequence, the greater the distance from randomness is expected to be.

This result (for the original ATCG upstream sequence) is therefore unexpected since it becomes more random approaching the TSS. This region has a higher density of regulatory sequence and so is considered more highly functional than the further upstream sequence. Since this region probably possesses a more specific language, this should be reflected in the distance from randomness. Also the very high distance from randomness of *intron1* (much higher than *exon1* and *coding1*) sequence is difficult to explain for the same reason.

3.4.2 Difference in distance from randomness for R/Y and W/S

General Meaning of the results in terms of structure and function

In order to make sense of these results it is useful to consider what is known about the function of the different portions of the upstream and the other genomic regions studied here. Indeed the aim is to better understand structure and function by analysing the sequence. The structure of the upstream DNA is of interest as is the way in which the cell reads and interprets this region through protein-DNA interactions.

When converting or translating the base sequence of the DNA into purines and pyrimidines (or weak and strong bases), the idea is to see their relative and comparative roles within the sequence and at different locations. Regarding regulatory regions it is interesting to find out whether the cell prefers to read or interpret the DNA sequence as purines and pyrimidines (or weak and strong). Also how does this vary across the sequence?

The four bases of the DNA are each unique, but this uniqueness may be viewed as being the result of overlapping properties. It could be that these two classes of the base properties have differing levels of significance within the DNA. This would depend on role and location of the sequence within the chromosome.

It may be, for example, that in the upstream regulatory regions (and protein-binding motifs), the base property of ring structure is more important than the property of hydrogen bonding capability in the double helix, or vice versa. It is seen in the results that W/S and R/Y show varying distance from randomness profiles to each other within identical DNA sequences.

In interpreting these results, an assumption is made that a greater distance from randomness confers a relatively greater level of importance for a particular class of base property. Therefore, if the DNA sequence becomes more distant from randomness, the arrangement of those residues is taken to be more significant in that region. For example, coding sequence in general is expected to be more distant from randomness than non-coding. This expected difference though is seen only for W/S but not with R/Y.

The distance from randomness values suggest that the arrangement of R/Y dinucleotides is of a relatively greater importance in the non-coding sequences than in the coding sequences. Also, within the non-coding regions, the R/Y sequences possess a higher distance from randomness than the equivalent W/S sequences. Why would this be? It could be that the R/Y arrangement is more important in these non-coding sequences due to a requirement for certain DNA helical structures and/or formation of chromatin and the binding of proteins e.g. for regulation. This may be the case because the R/Y arrangement has more of an effect on DNA structure than does W/S.

In contrast, the W/S sequence looks to be more important within the context of coding DNA than it is in non-coding DNA. In the coding sequence the gap between distance from randomness values of W/S and R/Y is greatly reduced in comparison to the non-coding sequence. Also W/S is more distant from randomness than R/Y in the coding sequences. This is the only genomic region (of those studied) where this is the case and is probably generally indicative of coding DNA.

This result is in line with work (Almirantis et al, 1997) on the clustered structure of purines and pyrimidines. They have shown that coding sequences present a close to random purine or pyrimidine distribution and in contrast non-coding sequences are not homogenous in this respect. Even very early work showed that the purine/pyrimidine versus weak/strong properties of coding and non-coding sequences are different (Blaisdell, 1983). The differences were seen in runs of monomers.

Why would the W/S arrangement be more important than R/Y in the coding sequence? This question is difficult to answer. It may have something to do with the way in which the genetic code operates. The arrangement of codons in the coding sequence is such that first, second and third bases in the sequence are of varying importance (Karlin et al, 1996). Also, it is known that codon choices are biased.

Clearly each base is important when the triplet code is read and interpreted during protein synthesis. However, it may be (for some unknown reason) that for this process the specific arrangement or order of W/S bases takes precedence. For instance one important factor may be that the third base in the codon determines the specific amino acid via selection of a weak/strong base rather than the purine/pyrimidine. Also, it may be that within the coding region the R/Y (dinucleotide) sequence arrangement is of a lesser importance because structural properties are of lesser consequence here (relative to the non-coding region).

Following the same line of logic; within the non-coding sequences the arrangement of R/Y is of greater significance than it is within coding sequences. It may be that within the non-coding sequences, and particularly close to the TSS (containing the promoter) and the intron sequence analysed here, there is some inherent property in the sequence that utilises R/Y to a greater extent. This may be in the form of the general structure of DNA. The specific way in which these purines and pyrimidines and their arrangement is significant is unknown.

R/Y and W/S differences in the upstream: could this be due to regulatory DNA?

The *upstream1* region being the most distant from randomness (with respect to R/Y) than further upstream segments suggests the increased significance of the R/Y arrangement in this region. Since this location contains more regulatory elements, it seems likely that differences observed between this region and the further upstream are due to these sequences. Alternatively, perhaps the R/Y residue arrangement is significant in the flanking regions of the regulatory sequences rather than in regulatory sequences themselves. Also, even if this effect is

not due specifically to regulatory protein binding motifs, flanking regions may also be important for regulation.

If indeed the R/Y arrangement is of relatively greater importance (and the W/S arrangement of lesser importance) in regulatory sequences than in non-regulatory sequences of the upstream, in what way would this be so? Since proteins bind to regulatory motifs it is reasonable to consider that this may be related to protein-DNA interactions.

The possible role of the R/Y and W/S arrangement in protein-DNA interactions

Research by Luscombe et al, 2001 has clarified the relative contributions of the different types of amino-acid with DNA interaction to the overall process of protein-DNA recognition. They found that 2/3 of protein-DNA interactions involved Van der Waals bonds. 1/6 were hydrogen bonds and 1/6 were water-mediated contacts. Of all of these types of interaction 2/3 were bonds with the sugar-phosphate backbone. These do not seem to be directly related to the base sequence of the DNA and therefore were deemed to be associated with indirect readout due to DNA structure.

From this description Van der Waals interactions are very important. However, the authors also explain that whilst these contacts contribute to almost 75% of the protein-DNA complex, they relate mostly to random docking interactions between the protein and DNA and are used to stabilise the complex.

For docking of the protein, which is the initial step, structure and conformation are essential for the general fit of protein and DNA. This may be said to be associated with indirect readout. However, it seems that Van der Waals interactions are mostly involved with docking. Probing on the other hand involves direct amino acid base contact. This means that regulatory sequences must accommodate for both steps to allow for correct protein-DNA interactions.

Could it then be that the R/Y base arrangement of the DNA is more important for the process of docking, whereas the W/S arrangement is more important for probing? Docking interactions are the more significant since this is the major contributor to protein-to-DNA binding. This may be reflected in the DNA sequence, in that the R/Y distance from randomness is increasingly higher than that of W/S in the TSS direction. In particular in the upstream (location closest to the TSS) the distance from randomness value for R/Y is more than four times higher than for W/S.

In what way and why would the R/Y sequence be more associated with docking and W/S more associated with probing? R/Y steps have a greater effect on DNA structure than W/S steps. In contrast to this, it could be that the potential for direct readout between protein side chains and the DNA is more dependent upon the W/S sequence.

Work by Lamoureux et al, 2004, has suggested that an amino acid side chain recognises two consecutive base-pairs in the DNA binding element. The 3' base they suggest is recognised via direct readout whereas the 5' base is recognised through indirect readout. For indirect readout it is the purine-pyrimidine step and its level of flexibility that aids recognition. This provides support for the idea that dinucleotides are important in protein-DNA recognition and more specifically that their purine/pyrimidine content may determine indirect readout.

It has been explained that the majority of protein-DNA interactions occur via indirect readout. Certain bases and base-pairs are involved in this readout. It may be the case that fewer base-pairs are involved in the direct readout process than in indirect readout. If the R/Y sequence has a greater effect on indirect readout and the W/S sequence on direct readout, this would explain why distance from randomness values are greater for R/Y than W/S. i.e. If fewer bases are involved in a specific process then overall a greater proportion would follow the random sequence profile. Also, it may be that for direct readout nearest neighbour effects are less important.

R/Y is more distant from randomness than W/S in the entire upstream sequence

It is important to note that the R/Y arrangement is more distant from randomness than the W/S throughout the entire 10Kb upstream and not only close to the TSS. Distance from randomness values for R/Y are approximately x1.8 higher within *upstream3*-to-*upstream10* than for W/S. This gap simply widens in the downstream direction. This suggests that in general (in the 10Kb upstream) the R/Y base arrangement is more important than W/S and this can not be solely attributed to regulatory regions. This is also true for other non-coding regions, namely the intron.

Early experiments by Zhurkin, 1983, showed that the alignment of nucleosomes on DNA followed a periodicity whereby RpY and YpR occur at intervals of five to six base pairs. This periodicity may be an example of the importance of the ordering of purines and pyrimidines within the sequence. The non-coding DNA may generally play more of a structural role within the chromosome. This explanation would be in line with the idea that the R/Y sequence is a greater determinant of structure.

3.4.3 The influence of individual R/Y and W/S steps on DNA structure

The effect of specific R/Y steps on DNA structure and possible role in regulatory DNA

Now the questions that remain are: Why particularly are RpR and YpY over-represented, whilst RpY and YpR under-represented? Also why do they become increasingly over-represented (RpR and YpY) or increasingly under-represented (RpY and YpR) towards the TSS? Whilst this issue was mentioned in the results shown in chapter 2, it may now be assessed iteratively and in more depth in light of the results and conclusions so far. Also here the sequence is viewed only as purines and pyrimidines.

The general under-representation of RpY and YpR in the entire upstream sequence (and the other genomic regions) is due to the need to avoid or at least suppress structural instability. Perhaps an even greater precision or specificity of structure (i.e. without multiple possible forms) is required towards the TSS due to the requirement to be read by the transcription machinery, a role which involves protein recognition and binding. Therefore further RpY and YpR suppression, and increased YpY and RpR enhancement may be important for regulatory sequences. If so, this observed balance may be needed for successful protein docking. This is because successful docking requires the correct structural propensity and the R/Y sequence of the DNA is a determinant of the relative rigidity/flexibility of the helix.

But why specifically are RpY and YpR more suppressed, and YpY and RpR more enhanced towards the start site region? What kind of dinucleotide steps do these generate and how may this specifically relate to protein docking? One example is that alternating purines and pyrimidines can form Z-DNA (Tiesman et al, 1990). It is thought that the conversion of B-DNA to Z-DNA may act as a genetic switch in regulation of gene expression (Sheridan et al, 2001).

It has been shown (Yagil, 2006) that DNA tracts composed of homo-purines or homo-pyrimidines are over-represented in various eukaryotic promoters including in the human. These would be composed of RpR/YpY dinucleotides. Their over-representation therefore supports the results seen in this project.

The decrease in RpY and YpR composition and their increased suppression towards the TSS shows a possible avoidance of the left handed Z-DNA. However, a poly(GC) sequence can also produce this Z-DNA structure. Also the other RY steps are increasingly suppressed towards the TSS, GpC is present at the random level but its frequency is increased towards the start site. Therefore there is a possibility that this dinucleotide would be responsible for Z-DNA formation within this location.

Dinucleotide steps and DNA structure

Dinucleotide step stacking data (El Hassan et al,1996) show that weak dinucleotides can produce a wide variety of DNA structures. These include; rigid A-form DNA (ApT), flexible A/B form (TpA), and rigid B-form (ApA and TpT). The data produced for these base stacking experiments in fact suggest that steps that contain only weak bases do not produce a particular class of DNA structure. Instead these steps produce structures that are more dependent on whether the constituent bases are purines or pyrimidines. This observation is also true for dinucleotides that are composed of one weak base and one strong base.

However dinucleotides that are composed of two strong bases (SpS) do fall into a class of their own regarding the roll and twist angles they produce and the resultant DNA structure if present repeatedly. This is because SpS is a bistable step so it is able to adopt two extreme (either high-slide or low-slide, but not intermediate) conformations. This affects the DNA structure.

It may be that flexible DNA allows for an increased capability of protein-DNA interaction (Feuerstein et al, 1990). For example, it is thought that flexible DNA can more easily bind histones. Reducing the flexibility of DNA and introducing more rigid DNA may reduce this type of protein binding in the upstream.

If the DNA were more rigid though how would the proteins involved in gene regulation bind the DNA? The answer to this may be increased bistability of the dinucleotide steps. In this case the DNA still possesses manoeuvrability so that it can adjust itself to the protein (by altering its conformation) but this adjustment may be far more limited and does not include multiple intermediary conformations as is characteristic of flexible DNA. This may be more discriminatory for protein binding. Perhaps then suppressing flexible DNA reduces general protein binding, whilst enhancing bistable DNA still permits specific regulatory protein binding.

When the upstream is regarded as a W/S sequence, there is a clear trend for the dinucleotide representation towards the TSS. SpS and WpW are less suppressed and WpS and SpW less enhanced. Whilst SpS has been discussed, the reason for this observation with the other dinucleotides is unknown. In the following sections though, the possible role of the W/S sequence in probing is expanded.

3.4.4 The relative effect of R/Y and W/S on direct readout

Distinguishing of base-pairs within the major and minor grooves

The H-bonding donor and acceptor patterns of the base-pairs may help to explain the way in which direct readout of the DNA sequences occurs. A view of these patterns within the context of the DNA sequence may provide a simple (albeit incomplete) picture for how DNA sequences may be distinguished. This is despite the fact that the readout mechanism is unknown.

Within the major groove each base-pair (C-G, G-C, T-A and A-T) seems to possess a unique H-bonding pattern. This is due to the asymmetry of the location of donors and acceptors since the purine bases in the base-pair contain two locations for potential H-bonding. In contrast, in the minor groove only C-G and A-T base-pairs may be distinguished from one another. Therefore C-G and G-C are not differentiated. This is also the case with A-T and T-A.

Purine/pyrimidine dinucleotides and weak/strong dinucleotide steps

It has been proposed in this project that the W/S sequence may be more important for direct readout than the R/Y sequence in protein-DNA binding. H-bonding between the protein and bases of the DNA sequence is essential for direct readout. Therefore H-bonding recognition patterns will be discussed in terms of W/S and R/Y dinucleotide steps and a comparison will be made.

Since the dinucleotide is the simplest motif in the DNA sequence, an analysis of these H-bonding patterns for dinucleotides illustrates the potential for the recognition code in the DNA sequence context. Furthermore it may be that dinucleotide steps are in and of themselves important for protein-DNA binding.

The question is; do W/S dinucleotides form a H-bonding recognition code within dinucleotide steps? Alternatively, is the recognition pattern for H-bonding distinguished through the purines and pyrimidines in the dinucleotide steps?

The minor groove will be discussed first since its base-pair H-bonding patterns are simpler. Here the H-bonding pattern of potential donors and acceptors are distinguished and determined via the W/S property of the bases and therefore it follows that the W/S (dinucleotide

steps) sequence determines the H-bonding patterns. R/Y bases are not distinguished. Accordingly, for the minor groove the W/S sequence determines to a greater extent direct readout.

The situation in the major groove is very different and requires further discussion. As already explained, here the purine and pyrimidine bases in the base-pair are distinguished for their H-bonding patterns.

Effectively though, the purine-pyrimidine distinguishing factor in the base-pairs is different for the A-T/ T-A base-pairs and the C-G / G-C base pairs. This is true in the following respect; the pattern of H-bond acceptors and donor are distinguished between A-T and T-A due to their relative distance from the phosphate backbone. In contrast, the C-G and G-C base-pairs are distinguished from each other due to this factor and also due to the asymmetry of the donor/acceptor pattern. This difference may be important for recognition.

There are important unanswered questions regarding the H-bond recognition pattern. For example, is the pattern read out along the base-pair affected by the location of donors and acceptors relative the phosphate backbone or is it just the order of donors and acceptors, or both? If both factors are important, then in the major groove the four base-pairs are all distinguished from one another. If only the acceptor/donor order is important and not the distance from the phosphate backbone, only the C-G and G-C base-pair are distinguishable from one another.

Another issue is that of roll, twist and slide values and their effect on the donor/acceptor recognition pattern. These vary along the different base steps of the helix. They also vary depending on the bending and twisting of the helix. If the recognition occurs along dinucleotides or over longer sequence stretches these variations may affect the readout pattern, since there is a shift in the location of base-pairs relative to one another.

If however, the H-bonding pattern were recognised by amino acids over the sequence in a stable manner, it seems unlikely that roll, twist and slide should greatly affect this type of recognition, i.e. the direct readout. This means that there would be some leeway regarding relative locations of the donors and acceptors in the recognition pattern across the dinucleotide step. In the same instance it may be that the location of the donors and acceptors relative to the phosphate backbone is of a lesser importance. These are unknown factors at present in the process of recognition of the base-pairs by amino acids.

Conclusion for potential recognition patterns

In conclusion, the minor groove recognition pattern is distinguished by the W/S property of the bases. This would therefore be true also across the dinucleotide step and indeed over longer sequence stretches. In the major groove the recognition pattern appears to be distinguished both via the R/Y and W/S sequence. This distinguishing of the base-pairs is less clear though in the major groove for the reasons given. However, even given this potential uncertainty, the C-G base-pair is distinguished from the G-C base-pair even with respect to the order of donors and acceptors. At present though the H-bond donor and acceptor patterns do not provide sufficient information for the process and patterns of H-bond recognition during direct readout.

3.4.5 Limitations of the dataset and the experiments

The change in dinucleotide composition and representation across the upstream sequence imply alterations in structure. The interpretation regarding structural changes across the upstream was qualitative and was based on work done on base-pair steps roll and slide angles (El Hassan et al, 1996). These angles were observed in crystallised DNA oligomers. The data is reliable in this sense (since it is high resolution X-ray crystal data) but may not be the exact *in vivo* situation. There has been some controversy as to whether the flexible or bistable structures can be adopted in vivo (Ringrose et al, 1999). Also, the interpretation was qualitative rather than quantitative and based on general averaged out results over 1000 base and 250 base windows rather than short local sequences. The limitations for this experiment regarding genomic sequence data collection and the use of the genomic signature are identical to those of the previous chapter.

3.4.6 The overall message and questions that arise

The distance from randomness experiment revealed a change in this (non-random) property across the 10Kb upstream sequence of the human gene. However, the observed results were not as expected since more highly functional sequences such as coding regions were closer to the random model and the sequence became closer to randomness in the downstream direction towards the start site. This was with respect to the original (ATCG) sequence.

In this project different sections of proximal sequences were analysed and a gradual change was seen in the genomic signature. Conclusions were based on an assumption that the more distant from randomness the region, the more highly functional it is. However, this assumption is problematic and the results have proven not so simple. If one region is more distant from randomness than another, this may not necessarily mean that it is more highly functional. In fact, the R/Y and W/S translations of the original ATCG sequence revealed that the distance from randomness trends were opposing towards the TSS, since the R/Y sequence became more distance from randomness whereas the W/S sequence became closer to randomness.

One of the conclusions drawn was that the R/Y sequence may be relatively more important closer to the TSS than in the further upstream portions. If this were the case, the expectation would be for the R/Y sequence to be relatively conserved. However, there is a problem with this idea. SNP data show that transversion substitutions are much more common in the promotor sequence than anywhere else (Guo et al, 2005). If the R/Y sequence arrangement would be more important in *upstream1* (than further upstream), why would transversions be more common in this promotor containing region? This is an important unanswered question. It is clear however, that there is a change across the upstream sequence in this genomic signature.

It is important to remember that the R/Y sequence is more distant from randomness than W/S across the entire 10Kb upstream. This implies that the R/Y sequence and its dinucleotides have a more profound effect in the upstream than the W/S sequence regardless of the position from the TSS. This general difference is probably due to the R/Y sequence having a greater influence on DNA structure and greater emphasis of structure in this region. The results suggest that the R/Y and W/S sequences do in fact possess different relative levels of importance in the different genomic regions. The specific nature of these differences though is more difficult to ascertain.

It has been proposed that these results were due to differences in the relative role of the W/S sequence and the R/Y sequence in determining DNA structure. This in regulatory regions may mean that the R/Y and W/S arrangement influence the processes of protein docking and probing in different ways. The suggestion of the different relative importance of the W/S and R/Y sequences in docking and probing has not been proven and this issue raises many questions, some of which are addressed in the later experiments of this project.

Conclusions for Protein-DNA Interactions

Based on the analysis carried out, the following issues may be considered:

• The R/Y sequence (of dinucleotides) has a greater influence over helical structure than the W/S sequence.

• The SpS dinucleotide is exceptional since it generates bistability.

• The W/S arrangement of bases (in dinucleotide steps) are a greater determinant of H-bonding patterns than the R/Y arrangement. This is true for the minor groove and may also be partially true for the major groove.

The results show that the R/Y sequence is more distant from randomness than the W/S sequence and that this gap widens towards the TSS. Also, flexible steps are suppressed and stiff steps enhanced in this direction, these being determined by the R/Y dinucleotide sequence. The R/Y sequence influences structural features to a greater extent than the W/S sequence. Indirect readout constitutes the majority of the protein-DNA interaction, and relies on suitable structural features of DNA for this to happen. DNA structure may therefore become more important towards the TSS in the location of regulatory sequences.

These observations have led to the suggestion that the R/Y sequence (of dinucleotides) likely affects the ability of proteins to dock onto the DNA and the process of indirect readout to a greater extent than the W/S sequence. It has also been proposed that the W/S sequence affects direct readout to a greater extent. This is supported by the H-bonding patterns within dinucleotide steps that are distinguished by their W/S arrangement within the minor groove.

SpS dinucleotides are exceptional in that they specifically affect structure. This set of dinucleotides is very prevalent in regulatory regions. It is these unique combined properties that probably make them key dinucleotides within regulatory sequence. In conclusion the W/S sequence is likely to be a greater determinant of direct readout and the R/Y sequence a greater determinant of indirect readout, with the exception of the SpS dinucleotides for which there are overlapping properties.

4. Upstream Sequence Similarity using a Patterns Analysis

4.1 Introduction

4.1.1 Sequence similarity and levels of functionality

Mechanisms for generating sequence divergence or convergence

Across different types of sequence or alternatively across a particular genomic region, DNA sequences may become increasingly divergent or convergent depending on two possible factors. These two mechanisms of sequence divergence/convergence are as follows:

1. Specific and similar types of substitution event.
2. Selection for particular structural or functional features.

The density of SNP's in the human genome is higher in coding regions than in non-coding regions (Subramanian et al, 2003, Cargill et al, 1999). Why would this be if coding sequences are in theory more highly conserved? Also, more SNPs are found in promoter regions than further upstream (Guo et al, 2005). SNP data which reflect substitution events suggest whether sequences are relatively convergent or divergent. These SNP results suggest, for example, that the promoter region is more divergent than the sequence found further upstream. This is counter-intuitive since the promoter (across many genes in any one organism) contains sequences that posses common features. For example, the TATA box and also other regulatory elements.

DNA has an inherent tendency to assemble in a particular way so as to form a stable structure. The assembly of the DNA double helix may result in a non-random sequence. This has been the subject of discussion in chapter 3. Additionally, different sequence types, such as coding and non-coding DNA may have a tendency to select for specialised structural motifs. If specialised helical structures are specifically required, for example for transcription regulation, regions such as the promoter may display and increase in sequence convergence when compared to DNA that is further upstream.

The reality in terms of the sequence properties of human DNA is probably a combination of these two factors conjoined, namely mutation events and a tendency to select for required structural features. Specified function (or higher level functionality) of the sequence

adds an additional level of complexity and a possible added deviation from the random model that was discussed chapter 3. This higher level functionality is likely to also result in changes in relative convergence or divergence of DNA sequence across different genomic locations such as the region close to the TSS and the further upstream sequence.

It has been evident from the sequence composition experiments of chapter 2 that compositional changes occur across the upstream sequence of the human gene. Also, dramatic compositional differences exist between the sequence of the promoter and the coding DNA. Therefore it is true to say that sequence properties change according to positional location in the genomic DNA, and this in turn depends on gene structure.

When nucleotide composition changes as it does across the upstream region of the human gene, a comparison of sequence similarity between the positional segments is expected to yield differences that correlate with those compositional changes. The greater the compositional differences between two locations, the lower the expected sequence similarity between them.

Changes in the level of sequence 'functionality' across the 10Kb upstream sequence

The issue of functional domains in the human genome has been the subject of the ENCODE Project (Birney et al, 2007; *ENCODE Project Consortium*). These are considered to be 'greater than gene-sized' functional domains. They were found to comprise 1% of the human genome and constitute a variety of functional domains. It has also been shown that ENCODE regions can be sub-divided into extended domains with common characteristics (see Thurman et al, 2007).

Conserved non-coding sequences are thought to correlate to locations of functional elements important for gene regulation (Hardison, 2000). If this is the case, it is expected that these conserved regions should be dense in regions of the genome such as the promoter and enhancer.

Regions within genomic DNA that contain functional domain and particularly locations with a high density of these may be said to be regions of high 'functionality'. In the upstream (or intergenic) region of the gene, for example, this would be locations containing regulatory elements as opposed to regions of low complexity or sequences containing a high density of repeats.

Intuitively a more highly functional region that contains the promoter and regulatory sequences would possess a higher sequence similarity across the set of human genes than regions of so-called lower functionality. The sequence motifs responsible for regulation would in theory confer a greater sequence similarity due to control modules and similarity of function.

Regulatory proteins bind to the 5' upstream region of genes via the recognition of specific sequence motifs. A regulatory protein may be involved in controlling the expression of more than one gene. These genes are then likely to be functionally associated. Gene clusters may operate in eukaryotes via the use of similar motifs upstream of different genes. Response elements identify genes that are under common regulation. A set of 5' upstream regulatory regions of different genes may contain similar sequence elements or motifs and differing combinations of motifs depending on the role of those genes. These can confer similarity between the upstream sequences of different human genes. These relative similarities in sequence would be dependent upon specific genomic location.

4.1.2 Relationship between sequence similarity and distance from randomness

The experiments of this chapter were designed as a continuation and follow-up of the distance from randomness experiments (of the previous chapter) wherein changes were seen across the 5' upstream sequence. It was also therefore of interest to compare sequence similarity across the upstream with the results for distance from randomness. In other words; what would the relationship be between changes in distance from randomness, sequence similarity trends and location (or functionality)?

Changes in sequence similarity across a genomic DNA region like the upstream are related to changes in distance from randomness across the same region. If a set of such sequences becomes more random in a particular direction, their level of similarity is expected to decrease. This is because increased randomness implies decreased functional association as is the case for decreased similarity.

4.1.3 The use of patterns in determining sequence similarity

There are different ways in which to test sequence similarity. Alignment is a very good method. However, this works well for a small number of sequences that possess high levels of similarity. In this case, there was a very large dataset (18,725) of upstream sequences, and therefore pattern similarity was a more appropriate method.

The problem of comparing large sequence sets

Why were common patterns used and not some other method such as sequence alignment? Sequence alignment is an obvious method for looking at sequence similarities. However, for large sets of data which need to be cross compared this method is too sensitive and less practical than a pattern comparison. This is because the alignment of sequences reveals only global similarities.

If for example, sequences are more distantly related, alignment is not very effective and such similarities are more likely to be picked up in a patterns search. This is because sequences that possess lower levels of similarity may possess relatively small common sequence stretches. In addition to this, these common sequences may not occur in a linear order. Pattern discovery or comparison tools therefore provide a simple, more effective and less time consuming (less computationally expensive) alternative.

The problem of cross-comparing multiple large sequence datasets

When comparing large sets of sequences to one another a system must be devised that is effective and a sampling technique should be employed that reflects the sequence sets. This is the essence of the problem: There is a set of 18,725 sequences; each is 1Kb in length, which is to be compared with four other sets of similar size. This amounts to a large quantity of sequence. This task would be computationally intensive and there is also the issue of effectively quantifying the data in a meaningful way. The advantages of using short patterns have already been highlighted. Also, specific patterns are not of interest, but rather the general pattern profile so that the sequence datasets may be compared in a relative sense.

4.1.4 The R/Y and W/S translated DNA sequence

In chapter3 results revealed that the upstream sequence becomes more distant from randomness towards the TSS, only when the sequence is viewed form the R/Y perspective, i.e. the R/Y translated upstream sequence. In contrast, for W/S translation (and also for the original ATCG sequence), the upstream sequence is closer to the random model toward the TSS. Thereby, R/Y translated sequence displayed the opposite trend across the 5' upstream to W/S translation (and the original ATCG sequence).

In the case of the 10Kb upstream region of the human gene analysed in the previous chapter the sequence displayed opposing distance from randomness trends depending on whether it would viewed as an R/Y translation or alternatively as a W/S translation. The reason for this (extensively discussed) was concluded to be due to an increased importance of structural features (stiffness versus flexibility) of the helix, which are primarily determined by the R/Y sequence. If the sequence similarity within the set of human upstream sequences is measured across separate 1Kb segments, it may be that the R/Y and W/S translated sequences would also display opposing trends towards the TSS.

4.1.5 Aims and experimental design

Large-scale comparisons of sequence similarities across the different 5' (1 Kb) upstream sequence segments were carried out. This was done both within each of the upstream sliding windows and between them. Since in the previous experiments, changes in nucleotide composition and representation were observed across the upstream, the next step was to analyse more specific sequence changes. This would be the next step to understanding the trends observed in the non-random characteristics seen across the upstream and how this may relate to functionality. The comparison of R/Y and W/S trends (i.e. for R/Y- and W/S-translated upstream sequences) was also carried through to this chapter since differences in these sequences were observed for distance form randomness.

For the sequence similarity experiments, instead of analysing 10Kb of the upstream sequence (as with the sequence composition experiments) only 5Kb was analysed. This is because in the previous experiments (chapter 3) distance from randomness changes were seen up to 5Kb upstream of the start site of transcription which, as already explained above is a related phenomenon.

Sequence similarity within different upstream locations

The changing trends in sequence similarity across the 5' upstream region would be useful for understanding the relationship between sequence divergence, functionality, and regulatory regions. In this experiment, the similarity of the set upstream sequences was studied to see if there would be any change from the 5'-end to the 3'-end. In other words, for the entire dataset (18,725) of genes, the sequence similarity was analysed within each of the different upstream positional segments (see figure 4.1).

The promoter region is dense with regulatory sequence and may be considered to be more highly functional than sequence that is further upstream (other than enhancers). This probably means that there is a bias towards the existence of certain motifs that are important for the process of transcription regulation. It is therefore relevant to find out whether this region has a greater sequence similarity within the set of different upstream sequences than the region that is further upstream.

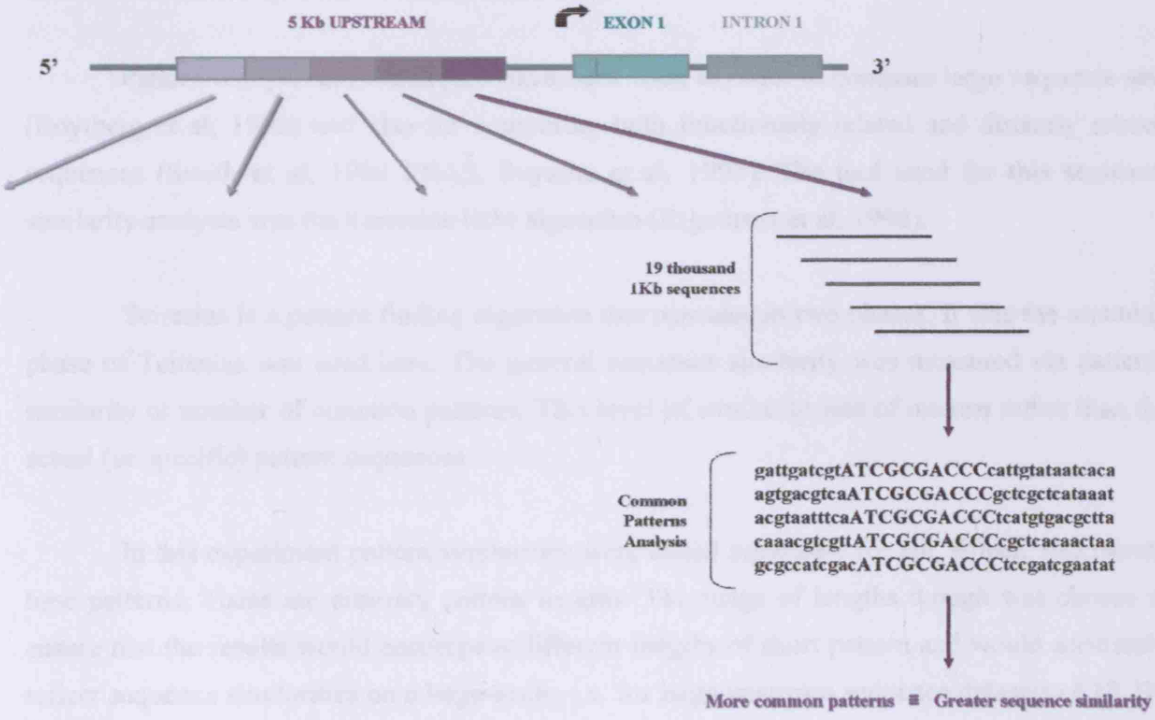


Figure 4.1: Diagram describing an experiment for the large-scale comparison of sequences within the different upstream positional locations.
For each of five 1Kb non-overlapping segments of the upstream region large-scale sequence comparisons were carried out. For example, for the upstream segment closest to

the start site, the entire dataset of 1Kb sequence strands was taken. A pattern matching program was then used to find common patterns occurring within the entire dataset. This procedure was repeated for each of the other four 1Kb upstream segments. Then any changes in sequence similarity (within each dataset) could be seen across the 5' upstream.

The expected result for this experiment would be that the positional segment closest to the TSS would contain the highest level of sequence similarity and that this similarity would decrease towards the intergenic region. This is because the start site region usually contains the promoter and possesses the highest density of regulatory motifs within the upstream. It is also thought to be the most highly functional sequence, as previously explained. Therefore the expectation is that the upstream becomes relatively convergent towards the TSS across the set of genes, so that sequence homogeneity would increase between adjacent segments further upstream.

Patterns analysis and the Teiresias Algorithm

Pattern comparison techniques have been used in order to compare large sequence sets (Roytberg et al, 1992) and also for comparing both functionally related and distantly related sequences (Smith et al, 1990 PNAS, Suyama et al, 1995). The tool used for this sequence similarity analysis was the Teiresias IBM algorithm (Rigoutsos et al, 1998).

Teiresias is a pattern finding algorithm that operates in two phases. It was the scanning phase of Teiresias was used here. The general sequence similarity was measured via patterns similarity or number of common patterns. This level of similarity was of interest rather than the actual (or specific) pattern sequences.

In this experiment pattern similarities were tested separately for ten, fifteen, and twenty base patterns. These are arbitrary pattern lengths. The range of lengths though was chosen to ensure that the results would encompass different lengths of short pattern and would accurately reflect sequence similarities on a large-scale, i.e. for large upstream sequence datasets of 18,725 1Kb fragments.

The output of the Teiresias program shows each pattern that is present in two or more of the upstream sequences. Each pattern is shown with its total number of occurrences in the entire dataset and the number of upstream sequences in which that pattern is found to be present.

 # Teiresias Output Sample:

Column1	Column2	Column3
2748	2230	GCTGGGATTACAGGCATCGT
1450	1000	GCCTCCCAAAGTGCTCTCCC
911	901	AAAGTGCTGGGATTAGGATA
580	560	CCTCCCAAAGTGCTGATCGT
181	135	CAAAGTGCTGGGATTCGCGC
10	10	CCAAAGTGCTGGGATCGATC
2	2	ATATCGATCTCCTAAATCGTA

First Column: Total number of times this pattern occurs in the set of input sequences
Second Column: Number of separate sequences in which this pattern occurred
Third Column: The pattern/motif sequence.

For the results the data from the second column was utilised. The number of sequences containing a particular pattern was used here rather than the total number of occurrences of that pattern. The reason for this was twofold. Firstly, a cross comparison of the whole set of upstream sequences was of interest rather than the total occurrence of a particular pattern. Secondly, it was favourable to avoid patterns that would be repeated, possibly many times in a small number of sequences, that would skew the results and make it appear as if the pattern was present in a large number of the upstream input sequences.

Needless to say, a large number of patterns would be common to any two upstream sequences, whereas a smaller number of patterns would be present say in one hundred upstream sequence fragments. The results would then show the common pattern occurrence within the entire dataset of 18,725 upstream sequences.

Sequence similarity between different upstream locations

The main issue that this section aims to address is regarding the relative uniqueness of the upstream sliding windows. This experiment was designed to ascertain whether the different locations of the upstream sequence possess different sequence motifs and therefore unique types of sequence.

A cross-comparison of each of the upstream positional segments was made with all of the other segments for sequence similarity. This was carried out by taking short patterns or motifs that are present in the 1Kb sequence adjacent to the TSS and finding their frequency and representation in the other upstream segments (see figure 4.2). This was done to see the difference between this region and the other upstream locations.

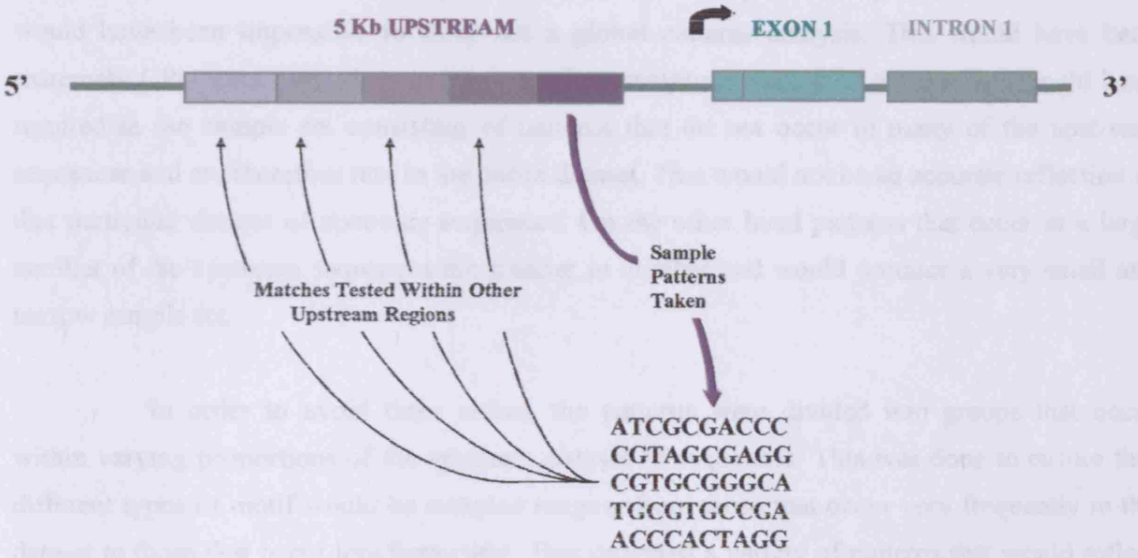


Figure 4.2: Diagram describing an experiment for the large-scale comparison of sequences across the different upstream locations.
The aim was to cross-compare the different 1Kb segments from the 5'-to-3' end, in order to find out if each is unique, to what extent and how this relative uniqueness changes. This was done by taking a sample of patterns from one location (for example, that closest to the start site) and comparing its frequency and representation with the other locations. This procedure was repeated for all the segments. The final picture enabled for a view of changes in sequence similarities and differences from one end of the upstream sequence to the other.

It was expected that there would be increased pattern uniqueness in the TSS direction. This is because towards the start site region there are boundaries such as the promoter and also the upstream sequence likely possesses sequences of lower complexity and lower functionality. Therefore the patterns were expected to be uniquely represented within that (near TSS) region. In other words, unique sequence features were expected closest to the start site due

to its role in gene regulation. It was expected that the further upstream the sequence being analysed the lower its level of uniqueness relative to surrounding sequences.

This process and its reasoning will now be explained. The patterns were divided across different occurrence groups. There are many different (for example, ten base) patterns that occur in only two (1Kb) upstream sequence fragments (out of the entire dataset of 18,725). However, only very few occur in one hundred of the upstream sequences.

Patterns had to be sampled from each of the upstream segment datasets, since it would have been impossible to carry out a global patterns analysis. This would have been extremely CPU intensive. A completely random sampling method of the patterns would have resulted in the sample set consisting of patterns that do not occur in many of the upstream sequences and are therefore rare in the entire dataset. This would not be an accurate reflection of that particular dataset of upstream sequences. On the other hand patterns that occur in a large number of the upstream sequences are smaller in number and would produce a very small and narrow sample set.

In order to avoid these issues, the patterns were divided into groups that occur within varying proportions of the upstream dataset of sequences. This was done to ensure that different types of motif would be sampled ranging from those that occur very frequently in the dataset to those that occur less frequently. This provided a variety of patterns that would reflect the entire dataset. Patterns were randomly sampled from each of these ranges or groups, resulting in a pseudo-randomised sampling method.

As well as the frequency of patterns, their representation was also worked out for each dataset. The representation of the pattern was considered to be the ratio of actual pattern matches as a proportion of the theoretically expected matches and may be considered comparable to the odds ratio calculation for dinucleotides (see section 2.1.6, chapter 2).

The representation (ρ_X) of $pattern_X$ is given as follows: $\rho_X = R_X / E_X$, where R_X is the real observed proportion *of* $pattern_X$ in the upstream sequence dataset and E_X is the random expect proportion, given the nucleotide composition of the pattern. The expected proportion (E_p) of a $pattern_X$ in the upstream sequence is given by: $E_x = (p_A)^{nA} . (p_T)^{nT} . (p_C)^{nC} . (p_G)^{nG}$; where p is the nucleotide proportion for the upstream sequences and n is the nucleotide frequency *in* $pattern_X$ (Xue et al, 2004). This calculates the expected proportion of a pattern in a single-stranded DNA sequence.

4.2 Methods

4.2.1 The upstream dataset

The DNA sequences for this project were obtained from the NCBI human genome database, build 35. The 10kb 5' upstream sequence of the gene was utilised. These human upstream DNA sequences were identical to those used in chapter2 (see methods section 2.2.1 for details). Five of the 1Kb upstream portions were used for the experiments described here; *upstream1*-to-*upstream5*, thereby spanning a length of 5Kb in total immediately upstream of the TSS.

4.2.2 Sequence similarity within different upstream locations

In this experiment, the sequence similarity within the different upstream positional segments was tested by finding common sequence patterns within each of the individual upstream datasets; *upstream1*-to-*upstream5*. For example, common patterns were determined between the 18,725 1Kb sequence fragments of the *upstream1* dataset. This would enable a view of changes in sequence similarity and relative divergence (or convergence) across the 5Kb upstream sequence.

The Patterns analysis and the Teiresias Algorithm

In this experiment pattern similarities were tested separately for ten, fifteen, and twenty base patterns. The scanning phase Teiresias, the IBM pattern finding algorithm was utilised to find common patterns separately within each of the positional upstream datasets; *upstream1*-to-*upstream5*. Teiresias scans the input set of sequences S, and finds patterns with support of at least K. This is the minimum number of common patterns that will be returned. The elementary pattern is one that is a <L, W> pattern and contains exactly L residues.

In order to clarify how the experiment was carried out it is useful to use an example. For the *upstream1* dataset of 1Kb sequences, tested for twenty base patterns the following parameters were used: The input set of sequences S=18,725. The minimum support K=2. In other words any patterns occurring in a minimum of 2 out of the dataset (of 18,725 1Kb)

upstream sequence fragments would be considered. The pattern length was fixed at 20, therefore L=20 and W=20.

The experimental output

The output of Teiresias showed each pattern that was present in two or more of the 18,725 1Kb upstream sequences for a particular dataset, such as *upstream1*. The output was then sorted to give the total number of common patterns present in x number of upstream sequences. I.e. the number of different shared patterns present in (x = *2,3,4,5,6, etc.....*) upstream sequences out of the total dataset which was tested (18,725 upstream sequences). The results would then show the common pattern occurrence within the entire dataset.

Logarithmic plots were produced for these results, separately for *upstream1*, *upstream2 upstream3*, *upstream4* and *upstream5*, i.e. the five datasets that were considered. This allowed for the relative intra-dataset pattern similarity to be examined. If a dataset possessed more common or similar patterns (a higher level of pattern similarity) than another dataset, this dataset was considered to possess a higher level of sequence similarity. These relative levels of pattern similarity were determined by considering the gradient and y-intercept values for each of the five plots.

Similar analysis carried out for R/Y -translated and W/S –translated upstream sequences

This common patterns study for each of the upstream 1Kb segments was also carried out on identical upstream sequence datasets that were translated into; (i) R/Y –translated sequences and (ii) W/S -translated sequences. The common patterns analysis was done in the same way using the Teiresias program, but this time only twenty base common patterns were searched.

4.2.3 Sequence similarity between the different upstream locations

In this experiment, the sequence similarity across the different upstream positional segments was tested by cross-comparing patterns derived from a particular region and finding matches in all other locations. Sample patterns from each of the sequence segment groups (*upstream1*-to-*upstream5*) were taken. The occurrence and representation of each pattern was then tested within that dataset, and then also within each of the other positional datasets. This allowed for an all-against-all comparison of the presence of patterns and therefore for a sequence comparison between the different locations.

The patterns/motif sampling method

Ten base patterns were used for this cross-upstream location analysis. The pattern sampling method was as follows: The output of the Teiresias patterns similarity results of the previous experiment (section 4.2.2, see above) was used. The output file contained a list of patterns together with the number of upstream sequences in which each pattern was present. For example, there would be such a list for ten base patterns occurring within the set of *upstream1* sequences. This list was sampled in a pseudo-random manner First of all the patterns were divided into groups which from which were subsequently taken a random sample.

The patterns to be sampled were first divided into groups that occur within varying numbers of the 18,725 upstream sequences. Ten different occurrence groups were formed in the following ranges:

Range1: $200 < x \leq 151$
Range2: $150 < x \leq 101$
Range3: $100 < x \leq 91$
Range4: $90 < x \leq 81$
Range5: $80 < x \leq 71$
Range6: $70 < x \leq 61$
Range7: $60 < x \leq 51$
Range8: $50 < x \leq 41$
Range9: $40 < x \leq 31$
Range10: $30 < x \leq 21$

The x value is the number of upstream sequences in which patterns occur. For example, range8 refers to patterns that were present in at least 40 and at most 49 different

upstream 1Kb sequences out of the dataset of 18,725. 50 patterns were sampled from each range, giving a total of 500 patterns sampled, from each of the five upstream datasets; *upstream1*-to-*upstream5*.

These range values refer to patterns present within x number of upstream sequences out of the whole dataset. Within each range, patterns were randomly sampled. Fifty patterns were taken from each of the above-given ranges. This brought the total of sampled patterns to five hundred, for each positional dataset. The random sampling was carried out using the C random number generating function *rand( )* (see appendix E.5 for more information). Each sampled pattern was 'replaced' after sampling.

Measuring pattern occurrence/frequency and processing the data

Now each of the patterns in the sample set taken from one upstream location was scanned for matches against each of the other upstream segment datasets. This was carried out using a pattern matching program, *COMMPATTS* (see appendix E.4 for more details). For example, each of the 500 sampled patterns taken from *upstream1* was scanned against all 18,725 1Kb sequence fragments of each of the other four datasets; *upstream2*-to-upstream5. An average value for frequency of pattern matches in an upstream segment was worked out for the five hundred patterns. This would be the average number of matches of a pattern sampled from one dataset (e.g. *upstream1*) and tested against another (e.g. *upstream2*).

Pattern representation and processing the data

As well as the frequency of patterns, their representation was also worked out for each individual upstream sequence dataset. For each pattern, the theoretical (or expected) frequency within the upstream was first calculated. The representation of the pattern was considered to be the ratio of actual pattern matches as a proportion of the theoretically expected matches. The expected frequency of pattern matches was calculated by considering the sequence composition of the pattern and that of the upstream dataset. This was done in the following way;

1. The upstream dataset sequence composition was worked out:

The A, T, C, and G nucleotide content of the upstream dataset being considered (e.g. *upstream1*) was taken. This was then given as a proportion value of the total number of nucleotides; p_A, p_T, p_C, and p_G respectively.

2. The pattern sequence composition:

For the (10 base pattern), the number of A's, T's, C's and G's was taken. These values are referred to as; n_A, n_T, n_C, and n_G.

3. The expect proportion (E_x) of pattern$_X$ in the upstream dataset:

$$E_x = (p_A)^{nA} \times (p_T)^{nT} \times (p_C)^{nC} \times (p_G)^{nG}$$

4. The representation (ρ_X) of pattern$_X$ was calculated as follows:

$$\rho_X = R_X / E_X$$

R_X is the real proportion of pattern$_X$ in the upstream sequence dataset, i.e. out of the total number of possible patterns and E_X is the random expect proportion, given the nucleotide composition of the pattern and the upstream dataset.

Pattern occurrence and representation for R/Y and W/S sequences:

The same upstream datasets (*upstream1*-to-*upstream5*) were used as described above. These sequences were translated into; 1. R/Y sequences and 2. W/S sequences. Therefore this yielded two new analogous upstream sets. The frequency and representation of patterns was tested within these two new upstream datasets in order to carry out the same cross-dataset sequence comparisons that have been described above for the original datasets.

There was however, a difference in the procedure. For these R/Y and W/S datasets the patterns used for the sequence match analysis were twenty bases in length. The patterns were taken from the Teiresias output of common patterns, from the Teiresias analysis of R/Y and W/S upstream sequence datasets. The random sampling of the patterns followed the same procedure as for the original (ATCG) datasets.

The same procedure of pattern scanning against the upstream dataset was carried out as for the original (ATCG) sequences and the identical system of data processing utilised. For the pattern representation within the upstream sequence sets the same calculation was used, only

this time there are two letters (or nucleotides) to consider instead of four. For example, for an R/Y twenty base pattern the calculation was as follows:

1. **The upstream dataset sequence composition was worked out:**
   The R and Y nucleotide content for all of the upstream dataset were worked out, where the proportion of R and Y is; $\mathbf{p_R}$ and $\mathbf{p_Y}$ respectively.

2. **The pattern sequence composition:**
   For the (20 base) pattern, the number of R and Y; $\mathbf{n_R}$ and $\mathbf{n_Y}$

3. **The expect proportion of the pattern in the upstream dataset:**
   $\mathbf{E_x = (p_R)^{nR} \times (p_Y)^{nY}}$

The representation value of the pattern was calculated as described for the original (ATCG) sequence sets. The sampling of patterns from each of the upstream positional datasets and the subsequent scanning of them for matches in all the other datasets (all-against-all dataset scan) was carried out as before.

4.3 Results

4.3.1 Similarity of sequences within different upstream locations

Common patterns for sequence viewed as comprising of ATCG

The results of this experiment show that there is a change in the profile of common patterns and therefore levels of sequence similarity within the upstream segments from the 5'-to-3' end. The *upstream1* dataset (which would usually contain the promoter) is distinct regarding the occurrence of common patterns (see figure 4.3). There are fewer common patterns in *upstream1* than in the rest of the upstream segments, *upstream2*-to-*upstream5*. The plots for *upstream2*-to-*upstream5* are fairly bunched together on the graph because they have a relatively similar profile of common patterns, although there is a slight increase in the number of common patterns from *upstream2*-to-*upstream5*, in the upstream direction.

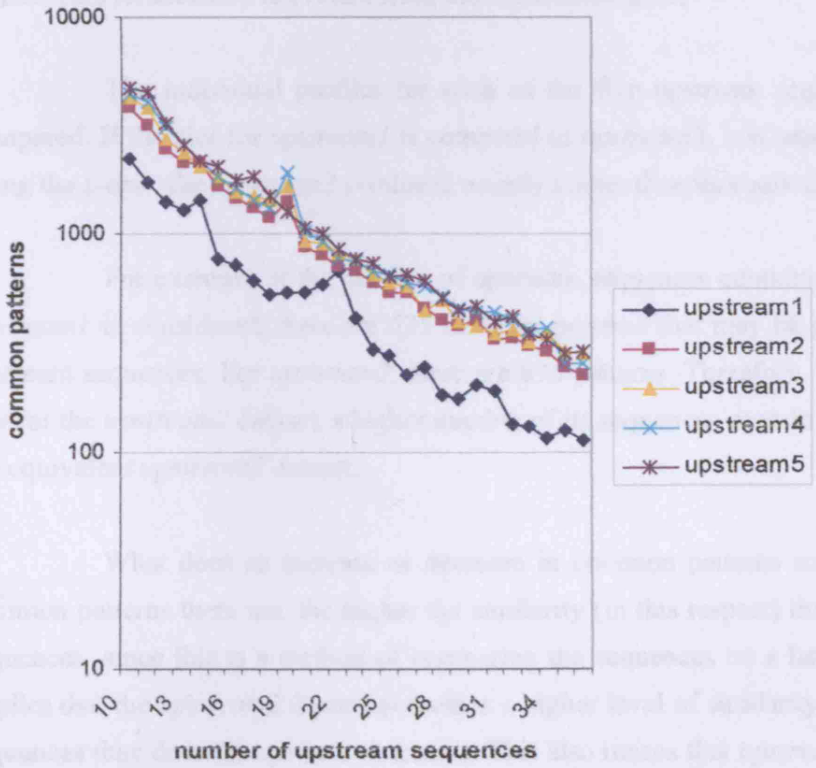


Figure 4.3: Graph presenting results of sequence similarity experiment.
This logarithmic plot shows common (20 base) pattern frequency in the upstream dataset. The number of common patterns was plotted against the number of upstream 1Kb sequence fragments (out of the whole dataset of ~18,725) that contain them.

As the number of common patterns being considered increases, the number of upstream sequences containing them decreases. This is the case with all five datasets.
When the five plots are compared, it is evident that *upstream1* is markedly different to the others. Here there are fewer upstream sequences (out of the entire dataset) that contain common patterns. Therefore with respect to 20 bases motifs, the set of *upstream1* sequences possesses the lowest level of sequence similarity in comparison to the other upstream segments.

The analysis of common twenty base patterns within the dataset of 18,725 1Kb upstream sequences shows that (for each of the segments, for example *upstream1*) as there is an increase in the number of common patterns considered there is a decrease in the number of upstream sequences containing them. Another way to explain this is to view the result as a profile of the number of upstream sequences containing identical patterns/words. Therefore there is a relationship between the numbers of upstream sequences that contain these identical words. This relationship is evident from the logarithmic plots.

The individual profiles for each of the five upstream segments were then cross-compared. If the plot for *upstream1* is compared to *upstream2*, it is seen that at identical values along the *x-axis*, the *upstream2* *y*-value is usually higher than its equivalent *upstream1* value.

For example, if the number of upstream sequences containing common patterns for *upstream1* is considered, there are 535 different patterns that may be present in any 20 of the upstream sequences. For *upstream2*, there are 875 patterns. Therefore, in general it can be said that for the *upstream2* dataset, a higher number of its sequences contain identical words than for the equivalent *upstream2* dataset.

What does an increase or decrease in common patterns actually mean? The more common patterns there are, the higher the similarity (in this respect) there is between the set of sequences, since this is a method of comparing the sequences on a large scale. Therefore this implies that the *upstream2* dataset possesses a higher level of similarity within its set of 18,725 sequences than does the *upstream1* dataset. This also means that *upstream5* sequences are more similar to each other than the sequences in the *upstream4* region, followed closely by

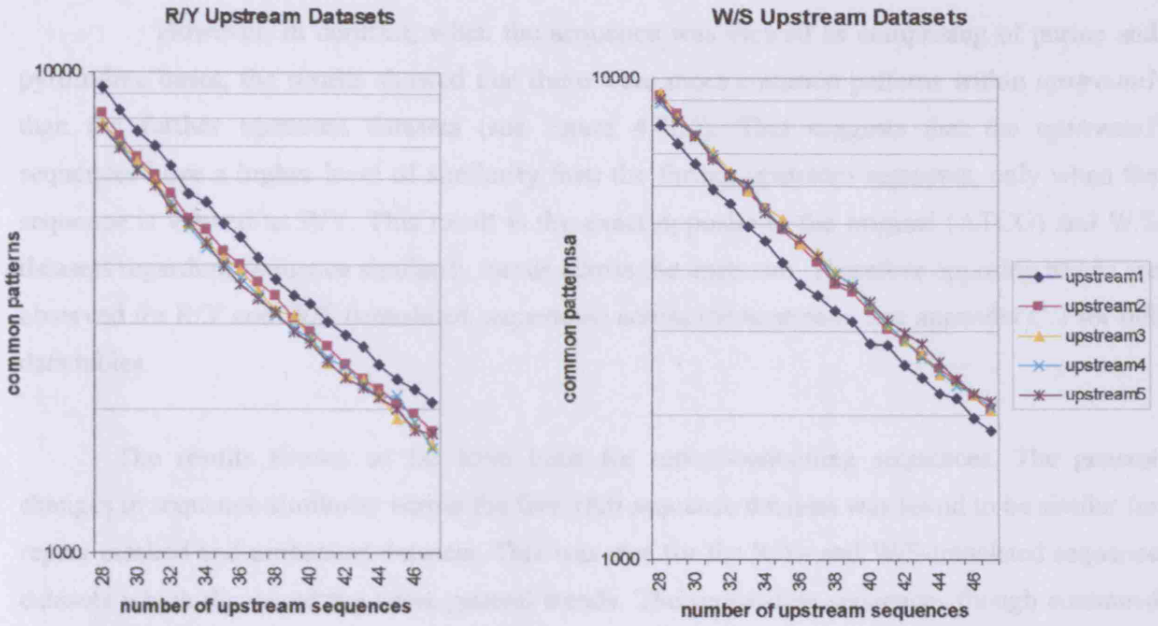
upstream3, and then *upstream2*. However these differences are far less pronounced than those between *upstream1* and the other upstream regions. The results described above are for twenty base common patterns. Similar trends were seen though for fifteen and ten base common (see Appendix C.1 for full details, graphs and tables).

The results of this experiment provide an insight into the sequence similarity within each of the sets of five different 1Kb portions of the 5Kb upstream region. It is possible to conclude that there is a trend of decreasing sequence similarity across the 5Kb upstream (towards the TSS), especially for *upstream1*.

Common patterns for sequence viewed as comprising of (or translated to) R/Y and W/S

The pattern similarity experiment, when carried out on upstream sequence datasets, converted to weak and strong bases, produced similar results to the original (ATCG) datasets (see figures 4.3 and 4.4). This was insofar as there are fewer common patterns in the *upstream1* region sequences than in the further upstream segments or windows (see figure 4.4(b)). Also, there was a slight decrease in common patterns within the datasets from *upstream5*-to-*upstream2* in the downstream orientation. However, *upstream1* was again the most distinctly different of the upstream locations.

This result suggests that within the dataset of *upstream1* sequences there is a lower level of general sequence similarity than the further upstream locations. Therefore when the sequence is viewed as consisting of W/S (translated) bases the profile is similar to that seen with the original sequence. This is not surprising, since why would it be any different?



4.4 (a) 4.4 (b)

Figure 4.4: Results of the sequence similarity experiment for upstream sequences that are viewed as; (a) R/Y –translated sequences, (b) W/S –translated sequences.

This logarithmic plot shows common (20 base) pattern frequency in the upstream dataset. The number of common patterns was plotted against the number of upstream 1Kb sequence fragments (out of the whole dataset of 18,725) that contain them.

This experiment and the details of the plots are the same as the graph above.

(a) R/Y graph:

There are more common patterns in the sequences of the *upstream1* dataset than for the further upstream datasets. Therefore the intra-dataset sequence similarity is highest for *upstream1*.

(b) W/S graph:

This result is similar to the result seen above for the original (ATCG) dataset (see figure 4.3), and is the opposite of that of the R/Y data. Here there are a lower number of common patterns within *upstream1* in comparison with the further upstream datasets.

In comparing the R/Y and W/S graphs it is obvious that there is an opposing trend of sequence similarity across the 5Kb upstream sequence.

However, in contrast, when the sequence was viewed as comprising of purine and pyrimidine bases, the results showed that there were more common patterns within *upstream1* than the further upstream datasets (see figure 4.4(a)). This suggests that the *upstream1* sequences have a higher level of similarity than the further upstream segments, only when the sequence is viewed as R/Y. This result is the exact opposite to the original (ATCG) and W/S datasets regarding sequence similarity trends across the upstream. Therefore opposing trends are observed for R/Y and W/S (translated sequences) across the upstream. See appendix C.2 for full data tables.

The results shown so far have been for repeat-containing sequences. The general changes in sequence similarity across the five 1Kb sequence datasets was found to be similar for repeat masked and unmasked datasets. This was true for the R/Y- and W/S-translated sequence datasets which displayed the same general trends. The repeat-free sequences though contained more noise. See appendix C.4 for full details of the R/Y and W/S results. The changes in pattern similarity for the original (ATCG) repeat-masked sequence proved too noisy to discern and clear differences between the five upstream sequence datasets (see appendix C.3).

4.3.2 Sequence similarity between the different upstream locations

Pattern comparison for ATCG sequences: Frequency

The results of this next set of experiments show the relative similarities (or differences) between the upstream segment datasets as opposed to within each segment. The results reveal that patterns derived from a particular region are present at a higher frequency in that region than in the other regions. For example, ten-base patterns derived from *upstream1* are present at the highest frequency in that region (i.e. their native region), and at a gradually decreasing frequency towards the furthest (*upstream5*) region. There are almost 45 pattern matches on average of *upstream1*-derived patterns within that dataset (see figure 4.5), followed by 35 patterns in the *upstream2* dataset, then 33 in *upstream3*, etc... This suggests that the datasets become more homogenous further upstream.

Patterns derived from *upstream2* are present at the highest frequency (45 patterns) in their native region. They are present at a lower frequency in *upstream3-upstream5* (41 patterns), and at a lower frequency still in *upstream1* (37 patterns). This result suggests that *upstream2* possesses distinct sequence features. Also it seems that *upstream2* is more similar to *upstream3*-to-*upstream5* than it is to *upstream1*.

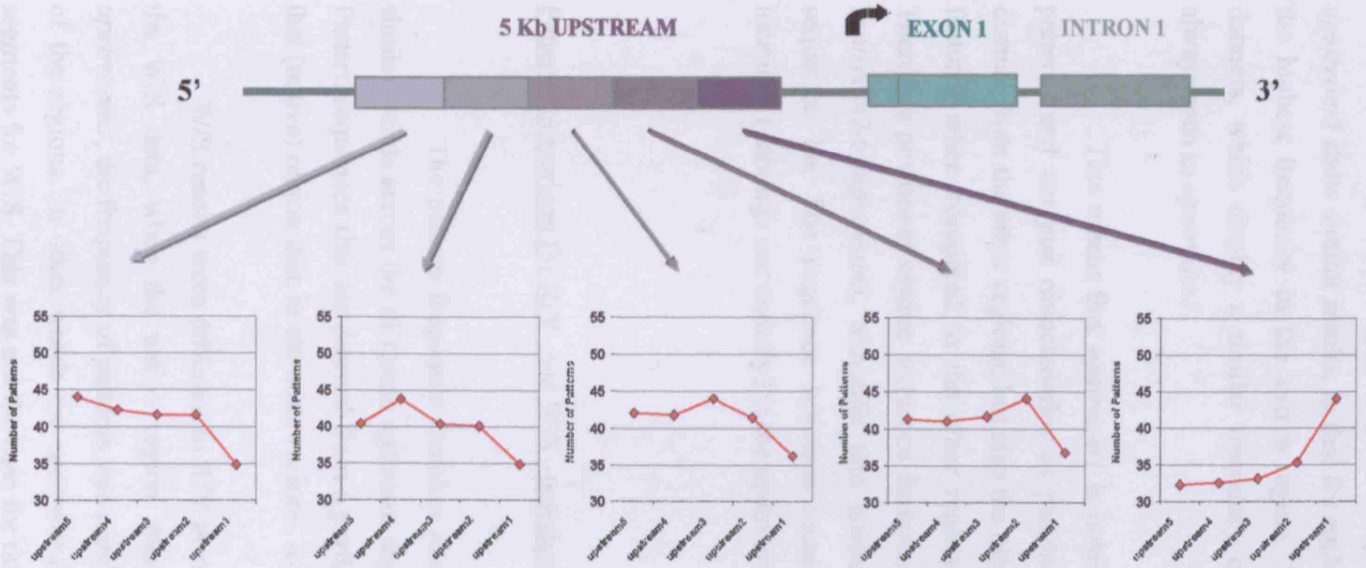


Figure 4.5

These are plots of pattern matches within each of the five upstream segments. The patterns were taken from each of the segments (denoted by the purple arrows), and the results of the pattern matches are shown in each graph (on the *y-axis*) for the patterns derived from that particular region against all of the other regions, including itself (the *x-axis*).

The graph on the far right, for example, shows the number of pattern matches (for the average pattern motif) within each of the upstream sequence datasets, *upstream1*-to-*upstream5*, for patterns that were derived from *upstream1*.

The number of matches is highest in *upstream1*, which is the native region, followed by *upstream2*, and then *upstream3*, *upstream4* and *upstream5*. Therefore *upstream1* is more similar to its neighbouring region *upstream2* than it is to the further upstream datasets and the datasets appear to become more homogenous further upstream.

In each of the graphs we see that the native region contains the highest number of pattern matches. Each of the upstream segments appears to possess some uniqueness with respect to its native patterns. We see also that *upstream1* appears the most 'different' in its pattern match score and that *upstream2*-to-*upstream5* segments are more similar to each other than they are to *upstream1*.

The pattern frequency plots for patterns derived from *upstream3*, *upstream4*, and *upstream5* show similar results, in that for each region the patterns derived from it are present at the highest frequency in the native region. This frequency is then lower in the remaining datasets, which display a similar frequency of patterns. The lowest frequency of patterns is always seen in *upstream1*.

This means that *upstream1* is distinct from the other regions with respect to larger patterns and not just dinucleotides as previously observed. It is not just *upstream1* that is distinct from the other regions, but also the other sequence portions have their unique sequence features when compared to the other regions, although to a lesser extent than *upstream1*. Therefore positional unique sequence features can be detected across the 5' upstream even in *upstream2*-to-*upstream5*, although this uniqueness is diminished the further upstream the sequence. I.e. The sequence becomes increasingly homogenous with respect to adjacent locations (although not entirely) in the upstream direction.

Pattern comparison for R/Y- and W/S –translated sequences: Frequency

The pattern frequency matches results for twenty base W/S or R/Y patterns show similar trends across the different upstream segments as for the ATCG results (see figure 4.6). Pattern sequences that are derived from a particular region are present at a higher frequency in that (native) region than in the other regions, as expected.

W/S results were different to R/Y primarily with regards to the *upstream1* region. For the W/S data, when the native region was either; *upstream2*, *upstream3*, *upstream4*, or *upstrream5*, the frequency of patterns was much lower in the *upstream1* than it was in any other of the regions. In other words the *upstream1* sequence was markedly different to the other segments for W/S. This was not the case for the R/Y data, whereby the frequency of *upstream1* patterns did not stand out in this way.

Clearly, the relative presence of W/S and R/Y (20 base) patterns are different; particularly in *upstream1*. The frequency plots for W/S and R/Y pattern matches show that the W/S plots are in general more similar to the ATCG pattern plots. This suggests that the W/S base arrangement affects the overall ATCG arrangement of bases more that the R/Y arrangement in this respect.

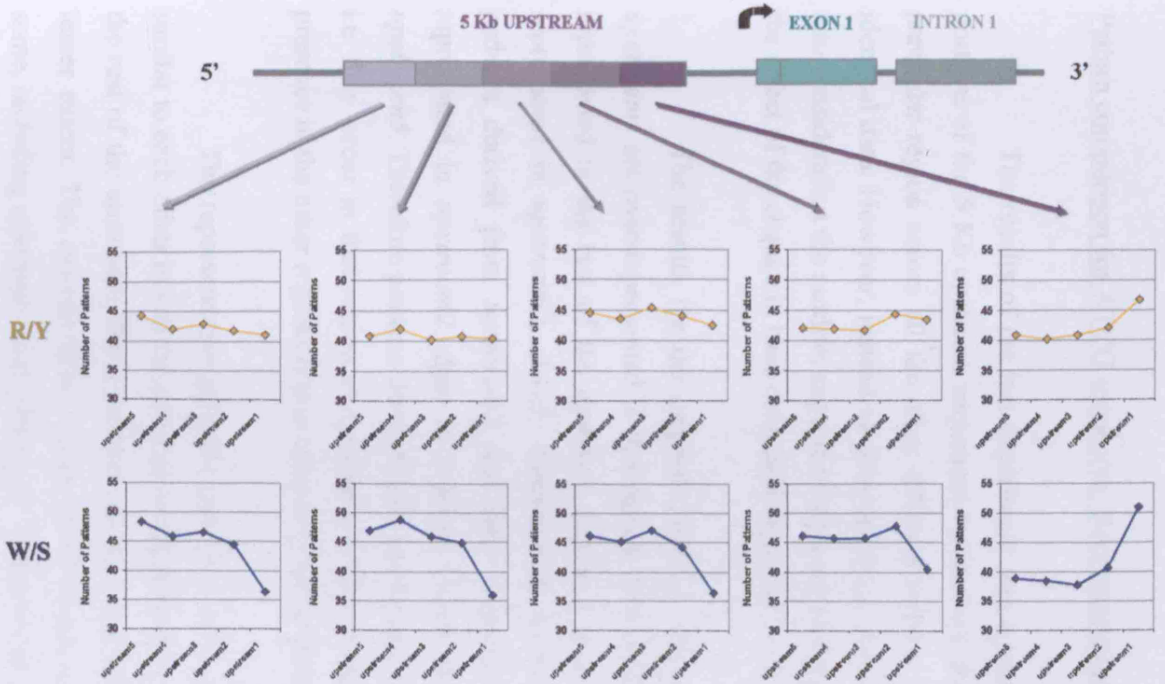


Figure 4.6:

These are plots of pattern matches within each of the five upstream segments. The patterns were taken from each of the segments (denoted by the purple arrows), and the results of the pattern matches are shown in each graph (on the *y-axis*) for the patterns derived from that particular region against all of the other regions, including itself (the *x-axis*).

R/Y results:

In each of the R/Y graphs we see that the native region contains the highest number of pattern matches. Therefore each of the upstream segments appears to possess some uniqueness in relation to the other segments.

W/S results:

The W/S results are in essence similar to R/Y. However, we see here that the pattern differences between the segments are more dramatic or accentuated those for R/Y. Also, for the W/S sequences *upstream1* seems to stand out from the other segments being the most different or distinct.

Pattern comparison for ATCG sequences: Representation

The results of the last experiment was a sequence comparison of the five different portions of the 5 Kb upstream sequences by showing the frequency of patterns derived from a particular region across all the other different regions. The following results are plots of the identical data. However, instead of showing frequencies, a ratio of patterns is given that takes into consideration the random expected frequency of the patterns. This eliminates or controls for the effect of the change in base composition across the upstream.

The results for the original (ATCG) sequences show that patterns derived from *upstream1* are over-represented in that region, as expected. These same patterns are also over-represented in the rest of the upstream, but to a lesser extent. The patterns are less over-represented in *upstream2*, closely followed by *upstream3-upstream5* (see figure 4.7). The patterns derived from *upstream2* also show a similar trend in that they are most over-represented in *upstream2*, then to a lesser extent in *upstream1*, followed by *upstream3-upstream5*. Therefore patterns derived from a particular region are most enhanced in that region, i.e. they occur in that region at a proportion that is higher than is randomly expected. Their presence in the other regions is also enhanced, but to a lesser extent.

The representation plots for patterns native to *upstream3*, *upstream4*, *upstream5* are similar to each other in that the patterns are slightly more enhanced in their native region than in the rest of the upstream. These patterns are also over-represented in the other regions, but to a lesser extent. The representation values in all of the non-native regions are more-or-less the same, including *upstream1* and *upstream2*. Therefore, in general these plots appear similar.

All in all this experiment shows that patterns that are derived from a particular segment of the upstream are more enhanced in that region than in any of the other upstream regions. This implies a certain measure of sequence uniqueness to each of the upstream portions analysed implying uniqueness of structure and function

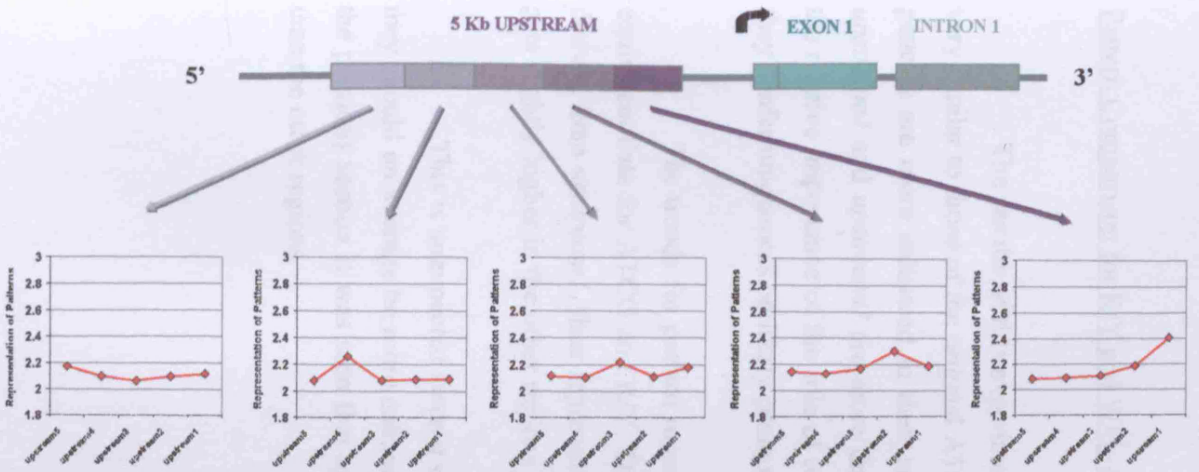


Figure 4.7

These graphs are representation plots of patterns taken from each of the five different 1Kb regions of the upstream sequence. The average representation of patterns (*y-axis*) from a particular region within all the other regions (*x-axis*) was plotted. Hence this shows a comparison of sequence derived from a particular upstream portion (e.g. *upstream1*) with all the other portions (*upstream2*-to-*upstream5*). The origin of the native patterns is indicated for each graph by a purple arrow.

The representation values take into consideration the composition and thereby changes in composition in the different upstream segments. This allows for a sequence comparison to be made whilst 'controlling' for changes in the nucleotide composition.

Each graph shows that patterns native to a particular region are relatively enhanced in that region. This is the case for all five upstream segments tested. Each of the five upstream segments possesses unique sequence features that are not simply the result of changes in sequence composition and in fact transcend the issue of changing composition across the upstream.

Pattern Comparison for R/Y and W/S –translated sequences: Representation

The results of this experiment show that representation trends of R/Y sequences are very similar to those of the original ATCG sequences (see figure 4.8). I.e. it is observed that the patterns are more enhanced in their native region than they are in the other regions. Again, *upstream1* and *upstream2* are more distinct than the further upstream portions. This confirms the relative importance of the role of R/Y patterns (and of course ATCG) in this region in that they confer uniqueness on likely promoter region.

The trends for pattern representation of the W/S sequence are very different to the equivalent data for ATCG and R/Y. This result is not in line with the expectation. For patterns derived from *upstream1*, their representation of these patterns is lowest in the native region and gets slightly higher in the other regions progressively towards *upstream5*.

This is unexpected since it seems that if patterns are derived from a particular region they should on average be more enhanced in that region. This is also counter-intuitive, since in the previous section it was seen that *upstream1* possessed a higher frequency of W/S patterns than the other regions.

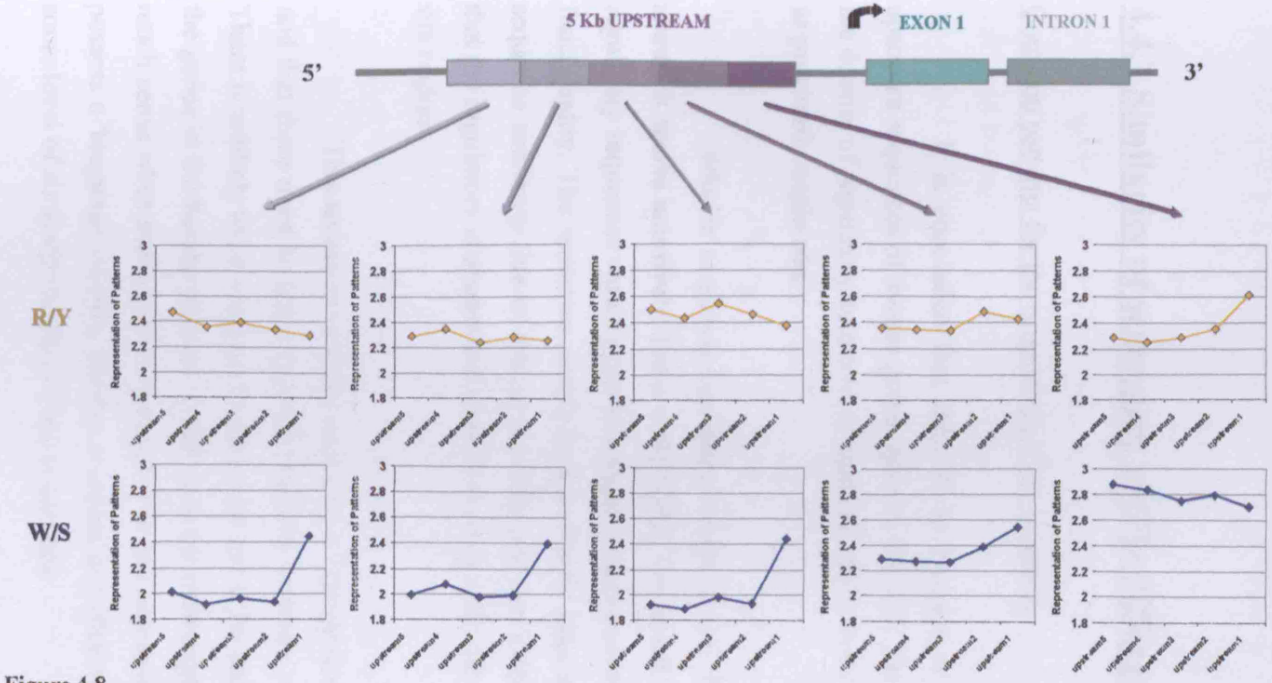


Figure 4.8

These graphs are representation plots of patterns taken from each of the five different 1Kb regions of the upstream sequence. The average representation of patterns (*y-axis*) from a particular region within all the other regions (*x-axis*) was plotted. These experiments and the layout of the results and graphs are the same as those seen in figure 5.5 of the last section, except that the experiments was carried out one equivalent R/Y and W/S upstream sequences.

R/Y result: These results are very similar to the result for the A/T/C/G experiment (see figure 5.5). Patterns derived from a particular region are relatively enhanced in that region, in all the upstream segments. This suggests that each of the five upstream segments possesses unique sequence features that are not solely due changes in sequence composition.

W/S result: These results are very different to those seen for R/Y (and A/T/C/G), which is evident in the graphs. If we ignore the *upstream1* result, we see that for all the other segments, patterns that are derived from a particular region are more enhanced in its native region in comparison with the others. *Upstream1* though is completely different. Its native patterns are relatively suppressed compared with all other regions. Also, patterns from all other regions are relatively enhanced in *upstream1*. This relative representation of patterns for the *upstream1* W/S sequence dataset is the exact opposite of that observed for all other upstream datasets.

4.4 Conclusions & Discussion

4.4.1 Similarity of sequences within different upstream locations

Common patterns for the original (ATCG) sequence

It is concluded that there is an increase in sequence divergence for the set of upstream sequences of human genes towards the TSS. This is unexpected since towards the TSS the density of regulatory sequence is high and also it is thought to be the more highly functional, as previously explained.

Why is sequence similarity lowest in the TSS direction? How could this come about? It seems intuitively that a more highly functional region that contains the promoter and regulatory sequences would possess a higher sequence similarity than regions of so-called lower functionality. The sequence motifs responsible for regulation would in theory confer a greater sequence similarity due to control modules and similarity of function. Alternatively it may be that the regulatory elements and their diversity confer this apparent divergence towards the start site region.

This seems an unlikely explanation though since gene regulation occurs in networks and that there must be some type of similarity between regulatory elements in different genes. There is unlikely to be a unique transcription factor for each gene, since this would mean half of the genes in the human genome would code for transcription factors alone. This does not make much sense when many genes act together in cell processes. Also, the regulatory elements must possess a 'language' of sorts albeit a complex one (that is not well understood), which confers some level of similarity between these sequences.

Another interpretation is as follows; the greater sequence similarity further upstream (in the intergenic sequence) is perhaps due to the presence of common structural elements or elements that confer stability on the DNA double helix. It may be that there is more similarity further upstream because the structural requirement of DNA in general results in a preference for certain elements, which are common to these sequences. Regardless of this possible explanation, it is difficult to account for the fact that the upstream sequence becomes increasingly divergent towards the sequence that contains the promoter and has the highest density of regulatory sequences.

Common patterns for sequence viewed as comprising of R/Y and W/S

This R/Y result of increased sequence convergence (or similarity) would have been expected in the first place for the original (ATCG) sequence. It would be more understandable if all the sequence types showed the same tendency or profile since the R/Y and W/S (translated) sequences are in fact derived from the ATCG sequence. Based on the expectation that the upstream become more convergent towards the TSS due to the role in regulation, this would in theory hold true regardless of whether the sequence is viewed as R/Yor W/S.

These observations beg two obvious questions; 1. Why would there be a higher level of sequence similarity towards the TSS with respect to the R/Y-translated sequence and 2. Why would the exact opposite be seen for W/S-translated (and original ATCG)? It was in the first place expected that the TSS region would possess a higher level of similarity due to regulation, which in theory would require similar elements across the board. Networks of gene expression mean that similar elements are required in the upstream sequence of different genes.

One possible interpretation is that perhaps purines and pyrimidines and the motifs that they form are particularly important towards the TSS, because they confer a relatively high level of similarity between sequences of different genes at this location. I.e. cause an increased convergence of the sequence. This may be associated with regulation and may be due to either with the regulatory motifs or possibly the sequences flanking them. Why would the W/S sequence similarity profile be similar to that of ATCG? It may be that the over-riding influence over the actual ATCG sequence comes from W/S in this respect. This would explain why the ATCG profile is similar to that of W/S.

It is apparent that identical sets of sequence can show reciprocal trends across the upstream (from the 5' end to the 3' end) with respect to the two different types of base property, R/Y and W/S. The first question is; how is sequence similarity interpreted? Sequence similarity is generally regarded as an implication of closer relationship of structure and function. It seems reasonable that a more similar sequence set may be regarded as one that possesses a more specialized function or that is more highly functional. However, as already mentioned there seems to be a paradox here regarding the R/Y and W/S sequences. So the second question is; how is it possible to interpret that the same sequence displays seemingly opposing trends?

It could be that a higher level of similarity does not necessarily correlate with a more specialized or higher level function. Perhaps more specific function can in some cases cause a divergence of sequence. The problem in interpreting these results is that in fact there are opposing trends with respect to R/Y and W/S. This relationship is likely to be very important. It

seems reasonable to conclude that R/Y and W/S possess different levels of functional significance across the upstream depending on location.

4.4.2 Comparison to the distance from randomness analysis

The link between distance from randomness and sequence similarity

In chapter3 results revealed that the upstream sequence becomes more distant from randomness towards the TSS, only when the sequence is viewed from the R/Y perspective. In contrast, for W/S (and ATCG), the dinucleotides are closer to the random model toward the start site. Thereby, R/Y displayed the opposite trend across the 5' upstream to W/S (and ATCG).

When sequence similarity across the upstream segments is considered, an opposing trend between R/Y and W/S (and ATCG) also exists. How are these two R/Y and W/S opposing trends in distance from randomness and sequence similarity connected? Since R/Y sequence similarity increases and distance from randomness decreases towards the TSS, it would seem at face value that R/Y motifs (or the arrangement of these bases) are more important here than further upstream. The equivalent W/S (and ATCG) results show the opposite trend. Also, it is interesting that where an increase in sequence similarity is observed, the sequence also becomes less random. Where sequence similarity is decreased the sequence in that location is more random. This makes sense intuitively.

It is possible then to infer that the R/Y sequence is in some way more important, relative to W/S close towards the TSS. The implication is that the sequence can be more highly functional or possess a more specific function and that this may be reflected in the sequence with respect to one category of base property (R/Y) and not the other (W/S). Also, R/Y and W/S possess a different level of emphasis across the upstream.

In what way could the R/Y sequence be of particular importance (relative to W/S) close to the TSS? Since this region is relatively dense with regulatory sequence, it could be that the R/Y arrangement is more important here. The importance of the R/Y dinucleotides in determining structure has already been discussed in chapter2 and may explain these observed trends. Since the R/Y sequence is a greater determinant of structure, perhaps the increased sequence similarity towards the start site relates to a need for greater structural similarity in this direction. This may be due to regulatory regions that require similar structural motifs in the DNA. If this R/Y similarity is due to regulatory regions and more specifically to protein binding

motifs, these structural elements may in turn be necessary for protein docking since this initial step of protein-DNA is very much dependent on DNA structure and flexure.

The W/S sequence on the other hand seems to be a source of greater sequence diversity close to the TSS than further upstream. This W/S influenced increased diversity may be related to regulatory sequences and may be important for their functioning. This relative diversity of W/S patterns may be related to probing by regulatory proteins which possibly require a diversity of sequence for this process to be discriminatory for individual proteins. It is important to note that the changes in R/Y and W/S divergence and convergence of sequence across the upstream are not necessarily due to protein binding regulatory motifs.

Functionality of genomic regions and SNP data

At this stage the issue of whether a higher level of sequence similarity and greater distance from randomness is associated with and implies higher or more specific functionality may be addressed. The density of SNP's in the human genome is higher in coding regions than in non-coding regions (Subramanian et al, 2003). Why would this be if coding sequences are in theory be more highly conserved? Also, more SNPs found in promoter region than further upstream. These observations seem counterintuitive. This is in line though with the increased divergence of sequence towards the start site for the ATCG original sequence datasets seen in this experiment.

Also, SNP data suggest that transversions increase towards the TSS (Guo et al, 2005). In fact their increase is greater than that of transition substitutions. This is counter to the observation in this work of increased R/Y sequence convergence towards the TSS. However, it is important to remember that these SNPs refer to comparisons between different individuals within the species which gives this data a different type of quality. This is very different to sequence comparisons within a genome.

Whilst transitions and transversions increase in the TSS direction for SNP data (comparisons between individuals) for the same gene, across the set of genes within the genome, comparisons show sequence divergence for the original (ATCG) sequence, but convergence for the R/Y sequence.

It is possible to conclude that higher level sequence functionality does not necessarily correlate with increased sequence convergence for comparisons across genes regions. Rather changes in these factors depend on how the sequence is viewed. From the

ATCG and W/S perspective here the change from the less functional to more functional location is opposite that observed for the R/Y sequence.

In this project it has been observed that sequence functionality does not to correlate in the expected way with distance from randomness values and sequence similarity across the datasets. Therefore the original stipulations made regarding these issues were clearly over-simplified. The situation is more complex.

4.4.3 Sequence similarity between the different upstream locations

There may be certain patterns (in the upstream) that constitute a background type sequence which is present throughout the upstream. In addition to these there may also be other types of pattern that are unique to a particular region and possibly relate to a structure and function that pertains more to that particular location.

This experiment was intended as a way to compare the sequences of the different positional upstream segments on a large scale. These sequences were arbitrarily divided into 1Kb portions. The results show that whilst there are many similarities there are also some differences that make each location unique, *upstream1* being the most unique.

The ATCG and R/Y datasets follow the intuitive expectation; that patterns derived from a particular region are present at a higher frequency and are also more enhanced in that region than in any other region. The W/S results are also in line with this general observation for regions; *upstream2*-to-*upstream5*. However, for *upstream1* the W/S sequences do not follow the expectation.

W/S sequence patterns native to *upstream1* are relatively suppressed in that region in comparison to the other regions. This is unusual since patterns are expected to be relatively enhanced in their native location. In order to address this issue it is necessary to break down the general question into parts.

Firstly, why would patterns native to a particular region be relatively suppressed in that region? It could be that these patterns (or their presence) must be more controlled for in the native region than in the other regions. In other words, suppression is favourable in such an instance despite the fact that the general trend is that patterns native to a region are relatively enhanced within that region.

It may be for example, that these patterns or sequences are very important to the native region as structural/functional elements, but their nature is such that it is better that they are scarce and the opposite would be detrimental in that region only and not in other locations. This may be true of gene regulatory elements, since enhancing them may be detrimental to the process of gene expression. Therefore the presence of these elements must be controlled so-to-speak.

Why are W/S patterns native to *upstream1* relatively suppressed in that region and patterns derived from the other regions relatively enhanced in the *upstream1*? It has been mentioned that relative suppression of native patterns is unusual and is probably due to the need to 'control' patterns in that region. It seems that there is something special about *upstream1* in that its native patterns must be relatively controlled or suppressed. Perhaps patterns that are native to the other positional segments do not need to be suppressed or controlled in *upstream1*. This makes *upstream1* a unique location in the 5' upstream sequence of the genes.

Initially it would seem that the uniqueness of patterns of the W/S *upstream1* sequence (relative to the other upstream segments) is due to a change in sequence composition in this region. The increase in strong and decrease in weak bases in the downstream direction appears to cause the different pattern frequency. The representation results though highlight an additional level of uniqueness.

The uniqueness of motifs in the upstream is not due solely to changes in sequence composition. The results show that for both W/S and R/Y sequences there are variations in sequence characteristics that cannot solely be attributed to changes in sequence composition.

When nucleotide composition changes as it does across the 5Kb upstream, a comparison of sequence similarity between the positional segments is expected to yield differences that correlate with compositional changes. The greater the compositional differences between two locations, the lower the expected sequence similarity. However, if this is controlled for, an additional level of information may be attained.

Differences in motif representation between the upstream positional segments imply that the formation of the sequence (motifs) is non-random and location specific. Therefore this difference in representation suggests unique structural and functional motifs. The changes across the upstream (especially between *upstream1* and *upstream2*) involve sequence arrangement and the formation of specific sequence. It seems that W/S motifs possess a special characteristic within *upstream1* relative to the further upstream datasets. The nature of these motifs though is unknown.

It has been observed (in chapter 2) that dinucleotide composition and representation varies across the upstream segments that were studied. In this current section, there has been additional support for the idea that there are changes across the upstream, in sequence characteristics that cannot be attributed solely to compositional changes. In this experiment these changes can be seen at the level of motifs. In addition, each of the 1Kb upstream segments possesses some of its own unique characteristics and the level of uniqueness diminishes the further upstream you look.

The motifs used here are obviously larger than dinucleotides and their analysis therefore present a more sensitive method for determining similarity and differences between sequences. In general these results suggest uniqueness of structure and function along the 5Kb upstream in these arbitrary 1Kb sequence positional segments. It is expected that the region containing the promoter would be different to other regions; however, here there are differences that span beyond the likely promoter location.

Since each 1Kb segment dataset of the upstream region that was analyzed possessed some level of difference in sequence to all of the other segments, there are likely to be both sequence, structural and functional gradients across this upstream region from the 5' to the 3' end.

In summary the work in this chapter confirms and extends the observation of the previous chapter. Firstly, it has been shown that there are increased sequence differences across the upstream towards the TSS. Secondly the observation that the R/Y and W/S sequences show opposing trends across the upstream is supported and is even more evident.

4.4.4 Limitations of the dataset and the experiments

Sequence similarity within the upstream positional segments:

The aim of this experiment was to analyse the relative sequence similarity within each upstream positional segment. The challenge here was that each one of these segments contained a very large dataset of sequences. An analysis of common patterns was used as a basis to identify similarities within the large dataset instead of alignments. Whilst alignment tools are very effective in measuring sequence similarity, a pattern analysis is also an accepted method.

There are however, disadvantages to this patterns analysis. Firstly, the results generated were qualitative and not quantitative. In this experiment the relative number of common patterns could be seen within each upstream dataset and therefore the relative similarity within the set could be determined.

Also, short patterns were used as a basis for analysis of sequence similarity, i.e. ten, fifteen, and twenty base patterns. Larger sequence stretches may in some ways be better. Shorter patterns are more likely to capture sequence similarities on a large scale, in a large dataset, whilst longer patterns would capture rarer similarities within the dataset. Only ten, fifteen, and twenty base common patterns were analysed here. This method, however, was simple and effective.

Cross-comparisons of sequence of the different upstream positional segments

This experiment posed an additional challenge to the previous one. Here the problem was one of comparing two large datasets of sequences to each other. Short patterns were also used here, however they had to be sampled from one dataset and their occurrence tested within another dataset.

There was a problem here with sampling patterns from the native dataset. The idea was that the sampled patterns should represent the dataset from which they were taken. A major limitation was whether the sampled patterns sufficiently reflected the native region. It is important to remember that this dataset is very large (18,725,000 bases) and that there are patterns within it that are very frequent and those that are very rare as well as a range between the two extremes. Therefore patterns were randomly sampled within different ranges. However, whether or not these actually represent the entire set is unknown.

A large set of five hundred patterns were sampled from the native upstream region. This is a good sample size although it was taken from different ranges. Of course a larger sample size of patterns would be more desirable. Patterns sampling from the native dataset of upstream sequences was only pseudo-random. The disadvantage of this is that the pattern sampling is biased. The advantage though is that the patterns better reflect the native dataset since these patterns are common to the dataset, since enriched patterns are selected for. This provides a more powerful tool for large-scale sequence comparison as was carried out here than a method of total random pattern sampling.

Whilst this sampling method has its problems it does provide a powerful sequence comparison methodology since it makes use of patterns/sequence motifs that are common to the

sequences within a set, then allowing for cross comparison across adjacent location sequences. Clearly there are limitations to this method; however a compromise had to be reached between generating an unbiased system whilst at the same time effectively representing a large dataset of sequences so that it could be compared to another large set.

In the cross-comparisons of sequence of the different upstream positional segments, the representation of sampled patterns was worked out. The method used for this has a limitation. In the calculation of the expect proportion (F_x) of pattern$_X$ in the upstream dataset ($F_x = (p_A)^{nA} \times (p_T)^{nT} \times (p_C)^{nC} \times (p_G)^{nG}$), the upstream dataset of 18,725 1Kb fragments is treated as if it is a continuous random sequence. This of course is not the case and leads to some error in the expect value.

4.4.5 The overall message and questions that arise

This result raises an important issue, which is that of sequence functionality and its relationship to convergence (or divergence) of an analogous set of sequences with respect to genomic location. First the issue of functionality and the assumptions made regarding the upstream sequence will be addressed. Secondly the subject of convergence (or divergence) in relation to this may be discussed.

The upstream sequence of the human gene contains regulatory sequence in the form of promoter and possibly enhancer which possess smaller units of regulatory sequence. Aside from these stretches of regulatory sequence the upstream contains spacer DNA of unknown function. It is known that the sequence contains nucleosome binding sites and may contain repeat stretches.

The promoter is usually located within 2Kb of the TSS and so this region may be considered the most highly functional (other than the enhancer), in terms of specificity of function across the dataset of genes. Therefore although it is true that non-regulatory sequence is largely an unknown entity, the assumption made in this work is that this promoter containing region is more highly functional.

Sequence divergence between different species can be used as a measure of the closeness of these species in an evolutionary sense. This can also be applied to related gene sequences within a species. From the divergence of a gene set it is possible to calculate for example, how long it has been since a duplication may have been generated among them.

With the upstream sequence though there are several differences. Firstly, non-coding sequence has been used, which is located along the upstream, according to its position. Therefore the related datasets have been grouped or divided in this way according to position from the start site. This is somewhat different to comparing homologous genes across species for relative divergence/convergence.

It is reasonable to expect that more highly and specific functional sequences should be more similar to each other. For instance, it could be that the more highly functional promoter containing region possesses a higher variety of sequence motifs which makes this sequence appear more divergent than the further upstream sequence. This phenomenon is possible although it would seem unlikely.

Whilst there are many unknown factors, experiments and their interpreted results must be based on theories that may possess inherent assumptions. The assumption here is that a greater similarity means a closer relationship or association between the sequences. The closer relationship would be due to specific function. If the idea is adhered to that the most highly functional region would be the one with highest sequence similarity (lowest divergence) and also that the promoter region is the most highly functional, the observed result seems contradictory. The result seems paradoxical.

This result showing increased sequence divergence towards the TSS though made sense only in light of the sequence becoming closer to randomness (for ATCG sequences) in the same direction. It seems logical that a more random sequence set should also be more divergent. The experiment that followed (the R/Y and W/S-translated sequence analysis) added an interesting new twist to the observations, and also strengthened what was seen for the (dinucleotide) distance from randomness analysis.

When the upstream positional segments were cross-compared for sequence similarity, it was found that each one possessed unique sequences. This result implies that these upstream 1Kb positional segments each vary in sequence and this probably means that this results in differences in structure also. This adds to the results seen in chapter2 whereby dinucleotide composition and representation suggest structural changes across upstream.

This uniqueness of sequence though is relatively greater towards the TSS. Overall the result shows that the upstream sequence is not a homogenous structure and that probably the sequence contains varying specificity along its length which is likely to be related to functional requirements. The specific differences between the positional segments and the result these differences in terms of structure and/or function are unknown. For the objectives set out in this

project general trends were of interest rather than the identification of specific sequence differences.

Larger motifs are a more sensitive method for detecting similarities and differences in sequences than are dinucleotides. In particular the increased R/Y-translated sequence similarity and distance from randomness towards the start site suggests a greater emphasis on structure (and structural similarity) in this direction.

Whether or not the changes in sequence similarity across the upstream can be attributed specifically to the regulatory (transcription factor binding site) sequences is still unknown. The differences in the R/Y and W/S-translated upstream sequence trends with respect to changes sequence similarity greatly supports the results of the previous chapter. The issue of the role of transcription factor binding sites is the subject of the next set of experiments.

5. Transcription Factor Binding Motifs: Avoidance of Random Binding of Regulatory Proteins

5.1 Introduction

5.1.1 Representation of the regulatory elements in the DNA sequence

DNA sequence motifs to which regulatory proteins bind tend to be up to twenty bases in length (Wingender et al, 2001). Regulatory elements are therefore made up of short and also specific sequences. This is true despite the fact that these sequences contain redundancies and that only some of the nucleotides within the motif may be specific. This means that in theory, equivalent sequence motifs (to those that bind regulatory protein) may occur 'by chance' in the 5' upstream non-coding regions of genes. As already mentioned, transcription factors are thought to bind any DNA sequence containing their target motif. Also, regulatory regions may span over very long DNA sequence stretches. Considering all of these factors, how do the regulatory proteins avoid binding to these sequences randomly or inappropriately, particularly in genomic regions where they exert their influence?

One possible answer is that regulatory elements are suppressed in order to prevent random binding. It could be for example, that these protein-binding motifs are present at a higher frequency in the upstream sequence than in the rest of the genome since this is the place in which they are required for their biological function. Also, their presence may be enhanced within specific locations of the upstream where they are required.

Regulatory motifs are thought to be over-represented in the promotor region (Long et al, 2004). Regulatory elements that act in other contexts, such as splice site elements are also found to be over-represented in those regions (Majewski et al, 2002). Therefore in general, it seems that regulatory motifs are over-represented in the regions in which they function.

A global analysis has been carried out of the distribution of transcriptional regulatory elements in human genomic functional elements (Zhang et al, 2007). This was carried out in ENCODE regions, within which the transcription regulatory elements were found to be present in clusters. In particular, these regulatory elements were found to be located in the region of known genes. Also they were enhanced near both transcription start and end sites. It has been suggested that the enhanced presence of the transcription regulatory elements near transcription end sites may be due to their function as regulatory elements (or promoters) for non-coding transcripts.

5.1.2 Mechanisms for avoidance of random binding of regulatory protein to the DNA

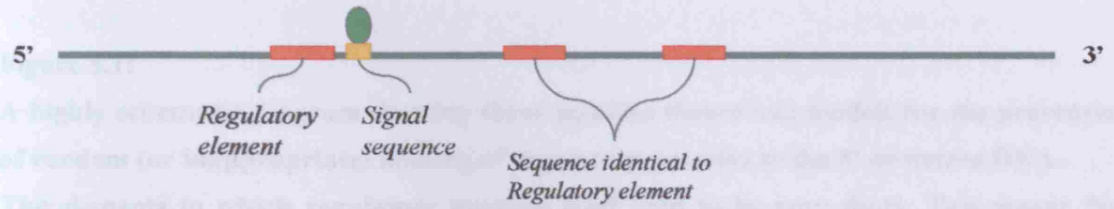
However, this story may not be so simple. For example, it may be that where the DNA is not exposed (due to chromatin formation etc...) the representation of these elements in any case is irrelevant. I.e. there would not be any binding, random or otherwise. The mechanism(s) of avoidance of inappropriate binding is unknown. Although in theory it may be explained in two simplified ways:

1. The first mechanism would involve marking the regulatory regions in some way so that they can be distinguished from other regions. This mechanism may include two general categories:
(i) The first is that there would be specific signal sequences in the upstream DNA to which proteins bind. These may differentiate between regulatory sequence and spacer sequence (see figure 5.1a). This dependence on another signal sequence greatly reduces the chances of non-specific binding. The problem with this idea is that no such signal has yet been identified despite much study of this region.
(ii) The second mechanism is that spacer sequence would be in some way protected from binding by regulatory proteins (see figure 5.1b). This may occur via 'masking' the non-regulatory regions. For example, the chromatin structure may produce this type of effect.
Both of these categories require the existence of specific motif sequences or signals within the DNA, albeit at unknown intervals and of an unknown nature. This introduces additional sequence features that may occur within (or surrounding) these distinct regulatory regions. Therefore spacer sequence and regulatory sequence are likely to contain distinct characteristics that are related to their different functions.

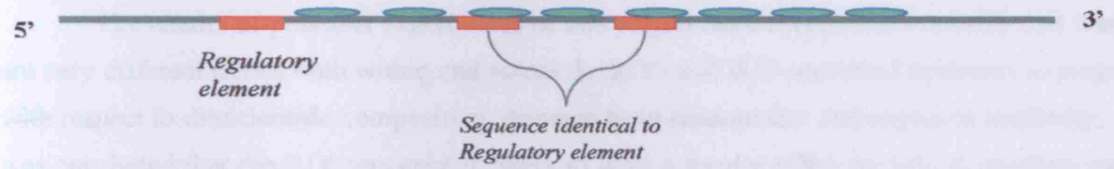
2. This would involve the representation of regulatory element motifs within the DNA sequence. By controlling the presence of the short sequences that are identical to those of the regulatory elements, it may be possible to avoid random binding of regulatory proteins in locations where this is detrimental (see figure 5.1c). Since regulatory proteins recognise and bind to specific elements on the DNA and these constitute short sequences, the proteins may in theory be able to bind to exposed DNA of identical sequence. If this were not a 'real' regulatory element, the situation would be at the very least wasteful and at most detrimental. Therefore, it could be that in such situations motifs identical to those of regulatory elements are suppressed in the DNA.

It is thought to be the case that the upstream transcription factors bind to any available sequence that is suitable. Therefore avoidance of inappropriate binding would be

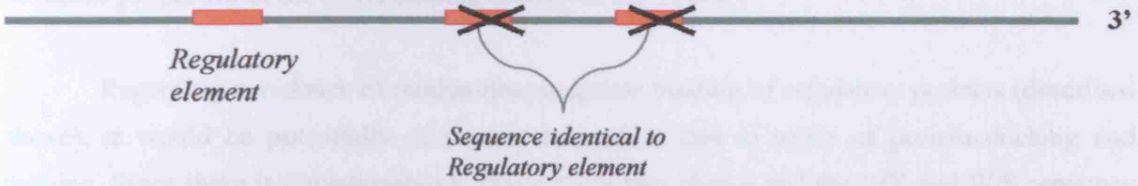
essential. In reality a combination of different mechanisms may be involved. Although the above-outlined models are likely over-simplified they are nevertheless useful.



a. This depicts the binding of a protein particle to a specific signal sequence (yellow box). In this model regulatory proteins can bind to the regulatory sequence only if this regulatory particle binds to its signal sequence first. Therefore it may be that regulatory DNA contains one or more of these signals that marks it as regulatory. This is one possible explanation for the prevention of inappropriate binding of regulatory proteins to the DNA. Alternatively, this signal may operate via a different mechanism than the binding of a protein particle.



b. Alternatively it may be that spacer sequence is in some way protected from the inappropriate binding of regulatory proteins whereas the regulatory element would be exposed. This is represented in the diagram by discs on the surface of the spacer DNA. The protection may be the result of DNA methylation, protein binding, nucleosome arrangement or some other conformational changes.



c. If there would be no signal and no spacer masking this third model may be relevant. Avoidance of inappropriate binding of proteins on the DNA (to sequences identical to the regulatory elements) would probably occur via a mechanism of the representation of those

elements. This may be described as a mechanism insofar as this process is a means by which the effect of random binding is avoided. In other words, in regions where inappropriate binding could occur, the regulatory element sequences may be suppressed. This means that they are unlikely to be present which would prevent such inappropriate binding.

Figure 5.1:
A highly schematic diagram showing three possible theoretical models for the prevention of random (or inappropriate) binding of regulatory proteins to the 5' upstream DNA.
The elements to which regulatory proteins bind tend to be very short. This means that these motifs may occur randomly in the DNA sequence. The issue is via which mechanism inappropriate binding of regulatory proteins to motifs that resemble their target is prevented.

5.1.3 Avoidance of random binding; the docking and probing steps

The results of previous experiments of this project have revealed collectively that there are very different trends both within and across the R/Y- and W/S-translated upstream sequence with respect to dinucleotide composition, distance from randomness and sequence similarity. It was concluded that the R/Y sequence is likely to have a greater effect on helical structure and that this is particularly important in the upstream DNA. It was furthermore determined that the R/Y sequence is likely to have a greater effect on the docking phase and indirect readout of protein-DNA binding and in contrast, that the W/S is likely to effect probing and direct readout to a greater extent.

Transcription factor binding sites (TFBS's) are motifs to which regulatory proteins bind, therefore the docking and probing steps of binding are of interest with regards to them. It may be that there is a relationship between these two phases of protein-DNA binding and the sequence properties of the DNA-binding motif, i.e. the TFBS.

Regarding avoidance of random/inappropriate binding of regulatory proteins (described above); it would be potentially of interest to analyse this in terms of protein docking and probing. Since there is a relationship between these two phases and the R/Y and W/S sequence properties, an analysis of these properties would be relevant. This would be in order to better understand how the TFBS sequence operates within the context of the upstream and genomic DNA sequence.

5.1.4 Aims and experimental design

The way in which regulatory motifs operate is clearly of great importance in the cell. Knowledge of the distribution of these motifs in the upstream is a step forward in gaining an understanding of how they regulate transcription. The aim of this experiment was to better understand this arrangement and also to investigate how inappropriate binding of regulatory proteins to the DNA may be avoided in the 5' upstream region of the human gene. This is pertinent since regulatory motifs tend to be short and may occur randomly within the genomic DNA sequence.

It is already known that DNA regulatory motifs tend to be over-represented within the genomic locations in where they function. The essence of the experiments of this chapter was to address the issue of avoidance of random binding of regulatory proteins to short sequences equivalent to those of the TFBS (in genomic DNA) specifically with respect to the docking and probing phases of protein-DNA binding.

Frequency of TFBS matches in the upstream region

This experiment was designed to show the location of the exact sequence motif that belongs to the TFBS across different positional locations along the 5' upstream region. Mismatches were not included in this analysis. This allowed for the distribution of sequence matches to TFBS to be seen across the 5' upstream positional segments. Specific single TFBS's were not of interest, but rather the average distribution. This is because the way in which TFBS's (as a group) operate rather than any specific motif is of interest.

A match for a regulatory protein binding motif would not necessarily define a location to which the regulatory protein actually binds, since there are more factors at play for this to take place. Indeed, protein binding may be dependent on context such as; location, chromatin structure, etc, and not just on the presence or occurrence of the correct motif sequence in the genomic DNA.

Another important issue is that of a possible boundary for regulatory sequences which was raised in previous experiments. In fact, in the previous experiments the issue of a potential boundary was raised only in so far as changes in sequence composition and sequence similarity that were seen across the upstream. The changing trends across the upstream segments plateaued in a way that was suggestive of a boundary for compositional properties.

The aim of this experiment was also to ascertain whether the regulatory element sequence distribution was in line with these boundary observations

This experiment was done to ascertain where in the upstream region the regulatory element matches would be most prominent. The TFBS's used were derived from the promoter region. Therefore the highest frequency was expected to be in the sequence closest to the TSS, since this is where most of the elements seem to have the greatest relevance in the biological context. The aim was to answer the following important questions; how are regulatory motif sequences distributed across the upstream? Are they present at a higher (or lower) level towards the TSS?

Representation of TFBS matches in the upstream region and genome-wide

The next and more important step was to see if regulatory motifs are over- or under-represented in the upstream sequence, within its different positional segments and also genome-wide. In other words; does the genome enhance or suppress the presence of the transcription factor binding motifs?

This was done in order to address the issue of inappropriate binding of regulatory proteins, previously described. Analysing the relative representation of the TFBS's would reveal some basic level information about the way in which binding of elements is avoided: either by (i) making the TFBS's sparse (suppressed) generally in genomic DNA, (ii) making the elements sparse in certain locations only, (iii) neither of the above, but instead via other mechanisms, for example, shielding the sequence e.g. chromatin formation. In this case the TFBS's (i.e. matches to them within the DNA sequence) would be present at the randomly expected level.

The results would provide clues about how the TFBS's operate and the issue of prevention of random binding would be addressed. The representation of a motif may in simple terms show one of three possible outcomes; 1. Representation at the random level, 2. Over-representation (enhancement), 3. Under-representation (suppression). The expectation was that the TFBS's be enhanced in regions where they are required and relatively suppressed in regions where they are not required (and also genome-wide), to avoid random binding events.

Since TFBS's consist of short sequences there is a possible risk of inappropriate binding of proteins, since identical motifs may occur in genomic DNA and in the upstream. How is this inappropriate binding avoided? The mechanism could involve suppression of short sequences identical to the regulatory binding motifs. Alternatively it could involve a different mechanism.

If within a particular genomic region (or within genomic DNA in general), there is a risk of inappropriate binding, the expectation is that these motifs be suppressed within that region Therefore any changes in motif frequency and representation across the upstream positional segments was studied as a part of this experiment.

Any differences when the sequence is translated into R/Y and W/S?

An essential dimension of this experiment was a continuation of the previously observed results (of other chapters of this project) regarding the importance of the two different groups of base characteristics; weak/strong (W/S) versus purine/pyrimidine (R/Y). Therefore the frequency and representation of TFBS matches within the upstream DNA was worked out for the TFBS and upstream sequences when translated to W/S, and also to R/Y.

This would allow for seeing the prominence of these characteristics of the bases within the regulatory elements in the context of the upstream sequence. The general question being; does the cell recognise the bases of these regions more in the form of the W/S property or the R/Y property? The objective was to reveal some information about how the cell reads or sees these regulatory elements within the upstream. More specifically this would permit the analysis of avoidance of random binding and occurrence of TFBS's in the context of the docking and probing phases of protein-DNA binding.

The dataset of TFBS's utilised for the analysis

The regulatory sequences for this experiment were taken from the TRED database (Zhao et al, 2005). This contains *cis* and *trans* acting factors including promoters and transcription factor binding sites. Motifs in this set are regulatory sequences originating from the promoter. Part of the database information included the location of these motifs within known promoters. For each motif there are also links to the relevant abstracts (in the Pubmed dataset) which describe details for that particular motif.

For this experiment TFBS's were utilised. The 1249 TRED elements ranged from ~4-90 bases (very few motifs were longer than this) in length and were represented across all the human chromosomes. Since the length of these regulatory sequences was over a very wide range with some sequences being relatively long, an effective method had to be

devised to test their occurrence and representation in the upstream datasets. TFBS's were scanned against the upstream sequence for exact matches.

The methodology used in this experiment had to take into consideration the length of the TFBS with respect to the total sequence (upstream) to be scanned for matches. The chance of finding at random a motif of length n within a stream of alphabet of z letters is given by; $\mathbf{P(m) = 1/z^n}$. For an (ATCG) DNA sequence;

Motif length 5; $\mathbf{P(m) = 9.8 \times 10^{-4}}$
Motif length 10; $\mathbf{P(m) = 9.5 \times 10^{-7}}$
Motif length 15; $\mathbf{P(m) = 9.3 \times 10^{-10}}$
Motif length 20; $\mathbf{P(m) = 9.1 \times 10^{-13}}$

The length of the upstream DNA sequence sampled for matches is 18,725,000 (1.8×10^7) bases in total. The sample space is even lower than this since the upstream sequence is not continuous. Since the likelihood of finding a motif at random decreases with increased motif length, the length of the TFBS motifs must be considered against the length of the entire upstream sequence space. The objective was to use as many of the motifs from the primary TRED dataset whilst at the same time not utilising motifs that are so long that they would not be reflected (or represented randomly) within the sample space. Therefore TFBS's were selected that ranged between 5-10 bases in length for the ATCG experiments. For the R/Y- and W/S-translated sequences, TFBS of length 10-20 bases were selected for the same reason.

There is also another issue regarding the longer (greater than 20 base) motifs. This relates to the fact that these do not necessarily reflect a discrete binding domain for regulatory proteins. Longer motifs probably contain stretches of intermediary sequences that are not involved in binding. I.e. a ninety base sequence (at the extreme end of the binding motif dataset) probably contains a collection of smaller motifs that specifically bind to proteins. Alternatively, the actual binding site(s) may occur over a narrower sequence space. In fact most binding motifs are thought to be up to twenty bases in length (Wingender et al, 2001).

5.2 Methods

Transcription factor binding motifs (TFBS) were taken and each of these motifs was individually scanned against different positional segments along the 5' upstream sequence of the human gene and also the entire genome for matches.

5.2.1 The upstream sequence dataset

The DNA sequences for this project were obtained from the NCBI human genome database, build 35. The 10kb 5' upstream sequence of the gene was utilised. These human upstream DNA sequences were identical to those used in chapter2 (see chapter 2, methods section 2.2.1 for details). Five of the 1Kb upstream portions were used for the experiments described here; *upstream1*-to-*upstream5*, thereby spanning a length of 5Kb in total immediately upstream of the TSS. Each of the upstream datasets consisted of 18,725 1Kb sequence fragments. Also the human genome-wide sequence was utilised in the same way as described in chapter 2, section 2.2.1.

5.2.2 The Transcription Factor Binding Motif Dataset

Origin of the binding motifs

The regulatory sequences for this experiment were taken from the TRED database (Zhao et al, 2005). This contains *cis* and *trans* acting factors including promoters and transcription factor binding sites. For this experiment (protein) TFBS's were taken from the database. These were 1249 motifs in total from the human dataset and they constituted the primary dataset from which motifs were selected for this experiment.

Length of the motif sequences

TFBS motifs were selected from the TRED database that ranged between 5-10 bases only and other elements were excluded. The removal of identical/duplicated motifs was then carried out, by testing the motif sequences against each other and redundant motifs were removed. This together with the selection described above left 154 motifs which were to be used as the test dataset of regulatory motifs.

5.2.3 Frequency of TFBS Matches (ATCG)

For each of the upstream positional segments; *upstream1*-to-*upstream5*, the frequency of matches of the TFBS's was individually determined. Each of the TFBS's from the sample set of 152 was taken and tested for exact sequence matches against a particular upstream dataset. The motif matches were scanned separately against the two DNA strands of the upstream sequence and the results for frequency of matches were recorded separately. The total number times that a particular sequence match was present an entire dataset of 18,725 1Kb upstream sequence fragments was recorded. This process was repeated for all 152 patterns. This was carried out using a program called *COMMPATTS* (see appendix E.4, for details).

The results were then processed. An average (median) was then worked out for the 152 patterns within each upstream dataset of sequences. This would be the average number of times that a binding motif (of length 5-10 nucleotides) is present or 'matched' in the entire *(e.g upstream1*) dataset. This average value would allow for the different upstream segments to be compared for the presence of TFBS matches. The process of calculating the number of matches for these 152 binding motifs was repeated for each of the other upstream datasets; *upstream2*-to-*upstream5*. A comparison of changes in motif distribution across the 5Kb upstream region could then be made.

5.2.4 Frequency TFBS Matches (R/Y- and W/S- translated sequence)

This TFBS searching experiment was also carried out for upstream datasets (identical to those described above) except that the DNA (ATCG -original) sequences this time were first translated into (i) R/Y TFBS sequences and (ii) W/S TFBS sequences. This therefore yielded two separate datasets of TFBS's. The primary TRED dataset was searched for sequences that ranged from 10-20 nucleotides in length. All other motifs either longer or shorter were excluded.

For each of these two TFBS datasets an all-against-all sequence comparison was made in order to eliminate duplicates/identical motifs. This left the total number of TFBS in the final datasets as follows; (i) The R/Y –translated dataset; 286 TFBS motifs, and (ii) the W/S – translated dataset; 224 motifs. The W/S dataset contained a lower number of motifs than the

R/Y dataset (despite being derived in an identical way from the same primary dataset) due to a higher number of duplicates in the W/S motif sequences.

The above-described scan for TFBS matches against the upstream sequences was repeated. However, this time the R/Y-translated dataset of 10-20 base TFBS motifs were scanned for matches against the equivalent R/Y- translated upstream and also R/Y- translated genome-wide DNA sequences. The W/S -translated 10-20 base TFBS were also tested for matches against the W/S –translated upstream sequences and genome-wide DNA, in the same way. The results were then processed as described above for the original (ATCG) dataset.

5.2.5 Representation of TFBS (ATCG)

The representation of each of the TFBS motif matches (e.g. for the original ATCG dataset of 152 TFBS's) was worked out for the *upstream1* sequence dataset. This involved using the real frequency of matches of the particular regulatory motif in question and then calculating it's expect frequency in the *upstream1* dataset.

Both the number of matches of that TFBS in *upstream1* and its expect frequency were considered within the entire dataset. In other words, the total frequency of matches in the 18,725 1Kb sequences was recorded (as opposed to the number of matches in each 1Kb sequence taken separately) and the expect value was also worked out for the entire upstream dataset. The expect frequency and the representation value were calculated as described in chapter 4, section 4.2.3.

The representation values were averaged out (using the median) across the datasets of 152 binding motifs, to give an average representation value of TFBS's (of length 5-10 nucleotides) in the *upstream1* dataset. This process was repeated for each of the upstream datasets; *upstream2*-to-*upstream5*. A comparison of TFBS representation across the different upstream segments could then be made.

5.2.6 Representation of TFBS (R/Y- and W/S- translated sequences)

The representation values for the TFBS were calculated using the same method described above for the;

(i) R/Y (translated sequence) dataset of 286 binding motifs, against the equivalent *upstream1*-to-*upstream5*, R/Y converted sequences.

(ii) W/S (translated sequence) dataset of 224 binding motifs, against the equivalent *upstream1*-to-*upstream5*, W/S converted sequences.

The expect values and the TFBS representation were worked out as described in chapter 4, section 4.2.3. The representation values were then averaged (median) across each dataset of TFBS for the R/Y -translated and W/S -translated separate sequence datasets. This provided an average representation value of TFBS (of length 10-20 nucleotides) in each dataset which could be compared across the 5' upstream positional segments.

5.2.7 Genome-wide representation of TFBS

Each of the TFBS's (e.g. of the original ATCG dataset of 152 TFBS's), was taken and scanned for sequence matches against the *whole genome* sequence. The frequency of matches was noted and then the expect frequency could be worked out for this particular motif in the *whole genome* sequence. This was done for each chromosome individually and then an average (mean) value taken across the chromosomes.

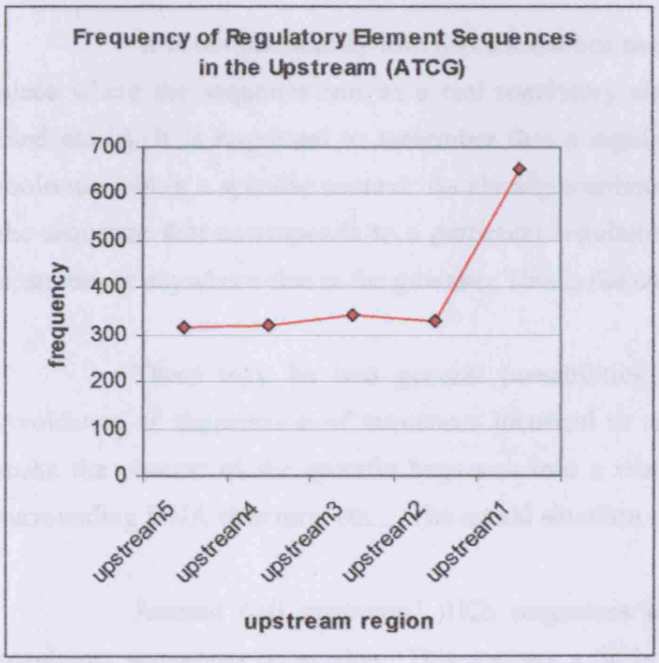
This motif frequency matching was carried out for each motif against each chromosome. Therefore the number of times that motif was present in all the contig sequences of that given chromosome was recorded. The expect frequency of that TFBS was also calculated within the same chromosome. The representation of the TFBS within that chromosome could then be calculated. The representation values for that particular motif in the different chromosomes were then averaged out to give the representation of the motif within the *whole genome*. This process was repeated for each of the 152 regulatory motifs.

An average (median) value of representation was then calculated across the entire set of 152 sequences to give the average representation of motifs of length 5-10 bases in the *whole genome*. This process was then repeated for the R/Y– and WS -translated datasets of TBFS's against an equivalent translated genome sequence.

5.3 Results & Conclusions:

5.3.1 Binding motifs (ATCG): Frequency

The results for the frequency of real regulatory sequences across the 5Kb upstream show that the frequency of the elements is highest within *upstream1*. An average regulatory motif sequence may be found within 651.5 (out of around nineteen thousand) *upstream1* sequences, that is approximately 3.4% of the dataset (see figure 5.2).



Average (median) frequency of regulatory motif matches in the ATCG upstream regions					
upstream region	*upstream5*	*upstream4*	*upstream3*	*upstream2*	*upstream1*
frequency	315.0	319.0	340.5	327.5	651.5

Figure 5.2: graph and data-table

Shown here is the number of regulatory motif matches in the entire dataset of (18,725, 1Kb) upstream sequences on the sense strand. The median number of matches for the set of regulatory motifs is given within each upstream positional segment. The results reveal that there are more than twice as many matches on average within *upstream1*, than each of the other locations (*upstream2*-to-*upstream5*). *Upstream2*-to-*upstream5* each possess a number of matches within a similar range: matches found between 315 to 340.5. This suggests that for these regulatory motifs, the main location for function is within *upstream1*.

This frequency is then reduced by more than 50% for *upstream2* (327.5 out of 18,725 or ~1.7% of the dataset), and remains more or less constant at the other locations. This is as expected, since the region of the upstream that is closest to the TSS and contains the promoter, is thought to possess the highest density of regulatory sequence. It is therefore not surprising that the sample of regulatory motifs generates more matches in this region compared to the further upstream datasets. However, the likely promoter location is within 2Kb upstream of the TSS and these motifs are most prevalent within 1Kb. The matches for the regulatory sequences within each of *uptsream2*-to-*upstream5* are at about the same level. The values range from around 300 to 350 (on average) motif matches. Therefore in this respect these regions are relatively homogenous.

It is not necessarily true that a sequence match for a regulatory element is actually a place where the sequence acts as a real regulatory element (to which the regulatory proteins bind etc...). It is important to remember that a regulatory element likely exists on the DNA molecule within a specific context. As already mentioned in the introduction, it is possible that the sequence that corresponds to a particular regulatory element could occur randomly in the upstream or anywhere else in the genome. This is the case because they are short sequences.

There may be two general possibilities for avoidance of random binding; 1. Avoidance or suppression of sequences identical to regulatory motifs. 2. Other methods that make the context of the specific sequence into a real regulatory sequence, such as position, surrounding DNA structure, etc... The actual situation is probably a combination of both.

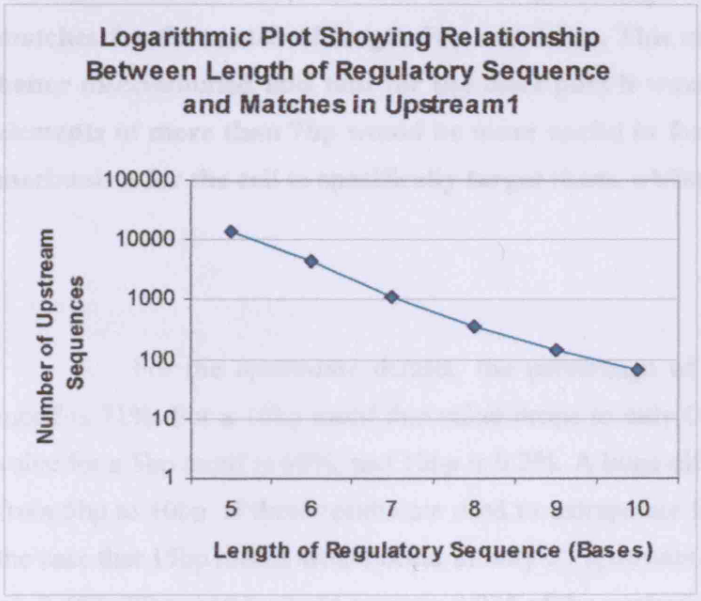
Around 650 *upstream1* 1Kb sequences out of 18,725 possess any one of the regulatory sequences on average. This appears a fairly high value if it were assumed that they were all real, since it would mean that about 3.4 % of the entire set of genes tested (on average) contains a regulatory motif. However, if they are not all real, how is inappropriate binding of the regulatory protein(s) that would attach to that sequence avoided? This question becomes more pertinent when it is considered that in a region such as the *upstream1*, the DNA is likely to be more exposed to such proteins during transcription.

Now the further upstream regions may be considered in light of these results. If it is assumed that the regulatory sequence matches to the upstream sequences are real, then around 350 genes possess an identical element sequence. What is more, this is the case in each of the 1Kb portions; *upstream2*-to-*upstream5*, which span up to 5Kb upstream of the TSS.

The results given so far describe data from an experiment that tests the number of upstream sequences (out of the whole dataset) that contain a regulatory motif match. 154 regulatory sequences were tested ranging from 5-to-10 bases in length. For a full breakdown of these results for different motif lengths see appendix D.1. The issue of whether the matches found for the regulatory sequences were in fact real has been raised above.

The occurrence of a particular motif within genomic DNA depends (among other things) on the length of that motif. In general the smaller the motif the higher the likelihood of its occurrence in the DNA sequence. Therefore smaller motifs in theory are more likely to be found randomly.

In order to address this issue a breakdown of the matches found for the different lengths of regulatory sequence has been presented (see figure 5.3). The number of specific bases involved in regulation is very relevant to this question, since the longer the regulatory sequence, the lower the likelihood of random occurrence. When the regulatory motifs were separated according to their length it could be seen that the longer the motif the less it would be encountered in the upstream. This follows a logarithmic relationship.



Regulatory Sequence Matches within *upstream1*

Length of regulatory sequence (bases)	Total Number of regulatory motif matches	Percentage of the dataset of 18,725 (1Kb) upstream sequences
5	13461	71%
6	4413	23%
7	1153	6.1%
8	354	1.9%

9	142	0.7%
10	65	0.3%

Regulatory Sequence Matches within *upstream2*

Length of regulatory sequence (bases)	Total Number of regulatory motif matches	Percentage of the dataset of 18,725 (1Kb) upstream sequences
5	13174	69%
6	3719	20%
7	903	4.8%
8	136	0.7%
9	142	0.7%
10	34	0.2%

Figure 5.3: Graph and data-tables:

The occurrence of a particular motif depends on the length of that motif. In general the smaller the motif the higher the likelihood of its presence in the DNA sequence. When the regulatory sequences were divided according to their length it could be seen that the longer the motif the less it would be encountered in the upstream sequence. This relationship is logarithmic (as shown in the graph).

A large difference can be seen in the percentage of upstream sequences containing matches for the motifs of length 5bp and 10bp. This means that larger sequences can be better discriminated and that for the most part it would seem intuitively that regulatory elements of more than 7bp would be more useful in for regulation, if there were no other mechanism for the cell to specifically target them, whilst excluding other sequences.

For the *upstream1* dataset, the percentage of upstream regions containing a 5bp motif is 71%. For a 10bp motif this value drops to only 0.3%. For the *upstream2* dataset; this value for a 5bp motif is 69%, and 10bp is 0.2%. A large difference in motif frequency is evident from 5bp to 10bp. If these results are used to extrapolate for motifs that are longer it would be the case that 15bp motifs would occur in only 53 *upstream1* sequences based in a gradient value of -0.463. 20bp motifs would occur in 0.025 of the upstream fragment sequences.

This means that larger sequences can be better discriminated and that for the most part it would seem intuitively that sequences of more than 7bp would be more useful for regulation as they are less likely to occur randomly. This though would depend on the type of regulatory sequence.

Some may be widespread among many genes whilst others would be present only within a small subset. For example, if a regulatory sequence would be needed only to regulate the activity of thirty genes in the genome it seems that 10bp or more of motif is required. This observation is of course not taking into consideration the representation of such an element, since it could be that a smaller element would in theory also be effective if its presence were suppressed in the sequence, or if some other mechanism were enacted to prevent this random binding.

There are some other important factors to be considered. It may be that the sequence involved in regulation is not a continuous stretch. This would mean, for example, that whilst in theory a 7bp sequence may be essential there would be gaps containing bases that are not essential. Within regulatory elements not all the bases are necessarily crucial for correct binding.

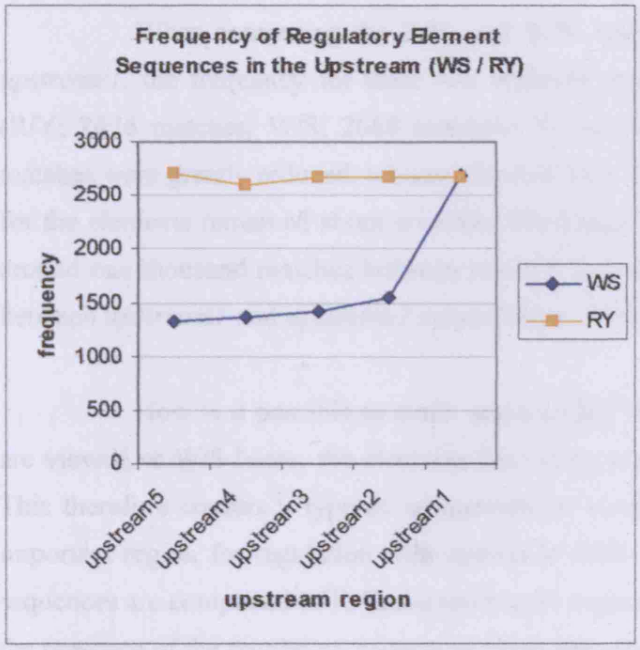
An extension of this idea is seen with composite elements which are two elements functioning together for gene regulation (Kel-Margoulis et al, 2000). Therefore it is not simply the length of the regulatory sequence that matters but also its context and probably its proximity to other elements. This adds a discriminating factor to these sequences and means that whilst for example, a >10bp regulatory sequence may be essential to discriminate thirty genes; this sequence may be split into two. Another way of looking at this is issue is that a shorter sequence may be utilized only if it acts in conjunction with another element.

5.3.2 Binding Motifs (R/Y- and W/S- translated): Frequency

This section contains results for the sub-division of the four bases into their two different classes of property; R/Y and W/S. The results are for the number of regulatory sequence matches within the datasets of different upstream regions (see figure 5.4) as was done for the previous experiment, only this time 10-20 base regulatory sequences were utilised.

When the upstream sequence is viewed as (translated to) W/S a similar trend is seen across the upstream sequence as was seen with the original ATCG data. There is a sharp drop in the number of matches between *upstream1* and *upstream2*. Unlike the original sequence though, there is a more gradual but only slight drop in the number of matches between *upstream2* and *upstream5*. In *upstream1* the frequency of regulatory sequence matches is 2688 (~14.1% of the whole dataset). This value falls by almost half in *upstream2*, to 1547.5 (~8.1% of the whole

dataset). From *upstream2*-to-*upstream5*, the number of matches is then slightly reduced until it reaches 1333.5 in *upstream5*.



Average (median) frequency of regulatory motif matches in the R/Y and W/S upstream sequences					
Upstream region	*upstream5*	*upstream4*	*upstream3*	*upstream2*	*upstream1*
Frequency (R/Y)	2702	2600	2672	2664	2676
Frequency (W/S)	1333.5	1364.5	1420.5	1547.5	2688

Figure 5.4: graph and data-table:

This graph and data table contain the results for regulatory motif matches in the upstream sliding window datasets on the sense strand.

The values on the graph and data-table are the number of regulatory motif matches within the entire upstream dataset. Regulatory sequences were tested (ranging from 10-20 bases in length) and the median value is given here, across the dataset of regulatory motifs. For the R/Y (translated sequence) data the number of matches is more or less the same from *upstream1*-to-*upstream5*. The actual numbers range from 2600-to-2702 (which means a match in approximately 14% of the entire upstream dataset).

The W/S (translated sequence) plot however, is very different. For *upstream1* the average number of matches is 2688. This value falls to almost half of this in *upstream2*, to 1547.5. From *upstream2-upstream5*, the number of matches is then slightly reduced until it reaches 1333.5 in *upstream5*.

In contrast, when the sequence is composed of R/Y, the frequency of motifs is about constant across the sliding windows. The frequency of matches range between 2600-to-2702, within each segment, which is approximately 14% of the upstream dataset.

When comparing the R/Y- and W/S- translated sequence results, it is seen that in *upstream1*, the frequency for these two different readings of the sequence is almost identical (R/Y; 2676 matches, W/S; 2688 matches). However, for the rest of the upstream, the W/S matches were greatly reduced, whereas for R/Y they were not reduced. Instead the R/Y matches for the elements remained about constant. Therefore, in *upstream2-upstream5* there is a gap of around one thousand matches between the R/Y and W/S data, this being due to the sharp drop between *upstream1* and *upstream2* values for the W/S data.

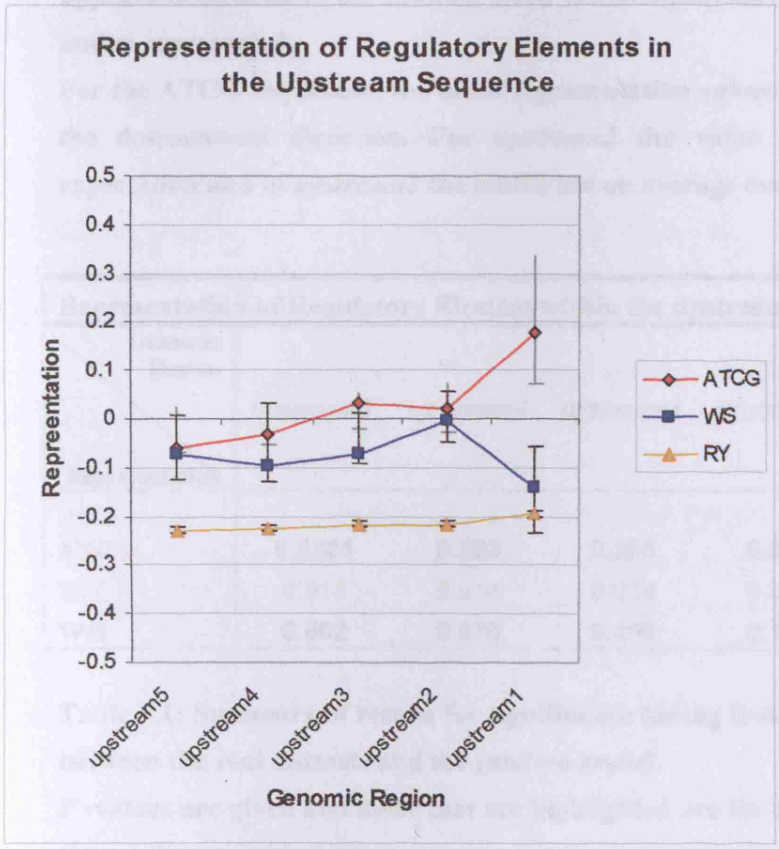
How is it possible to make sense of this? When the upstream sequences and motifs are viewed as W/S bases, the elements tend to be present at a higher frequency in *upstream1*. This therefore confers a type of uniqueness on *upstream1* and suggests that this is the most important region for regulation with respect to these W/S elements. However, when the same sequences are comprised of R/Y, the *upstream1* region no longer appears to be unique regarding the presence of the regulatory sequences, from this viewpoint.

Therefore, it would appear that the role of the regulatory elements across the upstream sequence is most important in a relative sense in the *upstream1* region for W/S sequences only. I.e. that the regulatory sequences are most prevalent in the *upstream1* reflects their increased presence and function in this region. This off course is assuming that the relatively high abundance of motifs tested reflects the presence of real, functional elements.

The results for the R/Y -translated sequences are much harder to interpret. It is clear when considering the set of regulatory motifs composed of R/Y that their presence and therefore their importance in the sequence is different to that of the W/S bases. According to these results the R/Y property of the bases has a different emphasis on the process of regulation than the W/S property. The nature of this is harder to understand. It may be though that the W/S sequence is more discriminatory for regulatory elements across the upstream sequence.

5.3.3 Representation of binding motifs in the upstream and genome-wide (ATCG)

The regulatory motifs were found to be over-represented in *upstream1* (see figure 5.5 and table 5.1). It could be that in the *upstream1* region, regulatory motifs are present at a higher frequency than is randomly expected since this is the region where they are most required for gene regulation. In contrast, in *upstream2* the representation of the elements is roughly at the random level which may be due to the fact that on the whole the regulatory elements are not required in this region.



Representation of Regulatory Element within the upstream and *whole genome* sequences

Genomic Region / Representation	*upstream5*	*upstream4*	*upstream3*	*upstream2*	*upstream1*	*whole_ genome*
ATCG	-0.060	-0.034	0.034	0.020	0.176	-0.006
R/Y	-0.229	-0.224	-0.217	-0.217	-0.195	-0.131
W/S	-0.074	-0.098	-0.072	-0.006	-0.141	-0.074

Figure 5.5: Graph and data-table

This graph and data-table show the average (median) representation values of regulatory motifs across the upstream sliding windows on the sense strand and also in the *whole genome*. A value of zero denotes representation at the random level.

The graph shows that the regulatory motifs are under-represented for R/Y- translated sequences throughout the upstream. The motifs become slightly less under-represented in the downstream direction.

Genome-wide these regulatory motifs are also under-represented, but less so than they are in the upstream (see the data-table).

For the W/S- translated sequences the regulatory motifs are generally under-represented although less so than R/Y - translated sequences. The W/S result shows a fluctuation of the motif representation values from *upstream5*-to-*upstream1*. For *upstream2* the result appears to be close to the random level. Within *upstream1* the motifs are on average more under-represented.

For the ATCG sequences, the motif representation values increase across the upstream in the downstream direction. For *upstream2* the value appears close to the random expectation and in *upstream1* the motifs are on average over-represented.

Representation of Regulatory Element within the upstream and *whole genome* sequences						
Genomic Region / Representation	*upstream5*	*upstream4*	*upstream3*	*upstream2*	*upstream1*	*whole_ genome*
ATCG	0.3831	0.363	0.365	0.301	0.023	0.382
R/Y	0.013	0.014	0.014	0.013	0.010	0.012
W/S	0.602	0.670	0.496	0.129	0.008	0.603

Table 5.1: Summary of results for significance testing (t-tests) of regulatory motif matches between the real datasets and the random model.

P -values are given and those that are highlighted are for datasets significantly different to the random expectation.

For the original ATCG sequences there is a significant difference between the actual frequency of regulatory motif matches and the random expectation for matches only in *upstream1*. For *upstream1*-to-*upstream5* and the *whole genome* sequence there is no significant difference between the real and random datasets.

The same result is evident for the occurrence of regulatory motif matches in the W/S-translated sequence datasets.

In contrast for R/Y –translated sequences, there is a significant difference between the real and random datasets in all the regions tested.

In *upstream4* and *upstream5*, the regulatory motifs appear under-represented, according to the average representation of motifs results. However, a statistical analysis of the actual frequency of matches versus the random expectation shows in fact that only in *upstream1* is there a significant difference between the real and random datasets. For all the other upstream positional segments there was no significant difference between real and random datasets.

The regulatory motifs are expected to be relatively over-represented in regions where they are needed. At the same time they are expected to be relatively suppressed in regions where they are not used or required, in order to prevent non-specific binding of proteins. Alternatively, in regions where the elements are not needed, their sequences may be present at the random level, if non-specific binding would not be an issue. This may occur, for instance if alternative mechanisms are in place.

The results of this experiment showed that on average the regulatory sequences that were tested were highly over-represented in *upstream1*; the representation value is 0.175. This means that on average the actual frequency of the regulatory motif matches is 17.5% higher than the expectation. For a full breakdown of these results for different motif lengths see appendix D.3.

The motif representation in the *whole genome* is on average at the random level. This was supported by the statistical analysis (see table 5.1) which showed no significant difference between the real and random datasets. See also appendix D.6 for a more detailed breakdown of the genome-wide results. The representation of these sequences genome-wide may be regarded as background level (or noise) since it is reasonable to assume that these elements are generally neither of a structural nor functional requirement genome-wide.

How can these motif representation results be interpreted? Since for *upstream2*-to-*upstream5*, the regulatory motif matches were close to the random level, it could be that suppression of these motifs is not used as a mechanism via which inappropriate binding of proteins is avoided. If the *upstream2*-to-*upstream5* matches had been under-represented, this would have suggested that inappropriate binding of proteins to the DNA is avoided by suppression of the sequences that correspond to the regulatory sequences.

This also seems to be the case in the genomic DNA in general in which the motifs were randomly represented. In a sense the entire genomic DNA sequence can be regarded as a control since the regulatory proteins do not possess a relevant function here. In other words the representation of regulatory motifs is not an issue here and therefore they are present at the random level.

5.3.4 Representation of binding motifs in the upstream and genome-wide (R/Y- & W/S- translated sequences)

When the sequences are viewed as W/S bases, the regulatory motif matches show that in *upstream1* they are under-represented. The representation value is -0.176 which means that on average the motif matches are 17.6% lower than the random expectation (see figure 5.5).

In *upstream2*-to-*upstream5*, the regulatory motifs are present at the random level. Although the motifs seem to be under-represented, particularly in *upstream4*, the statistical analysis has shown that only within *upstream1* is there a significant difference between real and random datasets. For *upstream2-to-upstream5* and the *whole genome* sequence, there was found to be no significant difference between the real and random datasets. Therefore, again the *upstream1* positional region stands out from the others.

The R/Y (translated sequence) regulatory motif matches reveal a very different representation profile. Here the motifs are far more under-represented or suppressed in all of the upstream segments than they are for the equivalent W/S- translated sequences. Within *upstream1* the R/Y regulatory motifs are under-represented, and the average representation value is -0.195. This means that the actual frequency of motif matches is 19.5% lower than the random expectation (see figure 5.5). The level of under-representation then increases only slightly and gradually further upstream, with the *upstream5* representation value at -0.230.

Within the R/Y- translated *whole genome* the regulatory motifs are also under-represented (representation value: -0.131). This effectively means that the R/Y motifs are generally suppressed everywhere, in all the tested regions. Also, the statistical analysis showed that there is a significant difference between the real frequency of motif matches and the random expectation throughout the upstream and genome-wide.

The motif representation trend across the 5Kb upstream sequence was found to be similar in repeat masked and non-repeat masked sequences. This is true for the R/Y- translated, the W/S-translated and also the original (ATCG) sequence. For a comparison with graphs and charts see appendix D7. Also the motif representation trend across the upstream was similar in the transcribed and non-transcribed strand, see appendix D.8.

5.4 Discussion

5.4.1 Binding motif frequency

Let us assume that a higher number of regulatory sequence matches in a specified region over other regions relates to the increased occurrence and actual use of regulatory elements. It could then be interpreted that *upstream1* is the most important positional region in the original (ATCG) sequence. This seems to be a reasonable assumption, since a higher presence is likely to relate to real regulatory sequence.

However, when this experiment is taken a step further, so as to see the arrangement of elements within the context of the two different properties of the bases (i.e. W/S and R/Y) two different parts of the entire picture are observed. When the sequence is translated into W/S bases, the result is in line with the observation made for the original (ATCG) bases. Therefore this property of the bases (the ability to be weak or strong) is probably the predominant property that exists within the sequence of these regulatory elements and it also discriminates them within the upstream sequence.

In contrast, the R/Y property displays a different level of importance and role in regulation. When viewed from this perspective, the prevalence of the TFBS motifs is the same across the 5Kb upstream. This indicates that from the R/Y perspective, *upstream1* is no more important as a region of regulation than the other upstream segments.

Another important factor to consider is the nucleotide composition of the regulatory elements. Weak and strong base composition changes dramatically across the upstream. If the base composition of the regulatory elements tends towards a bias for more strong bases and less weak bases, this would explain the higher frequency of matches within *upstream1*. However, it is important to note that if this is the case, it may nevertheless be of functional significance.

If the TFBS motifs are in fact high in CG and low in AT, this would explain the observation that their frequency becomes relatively higher in *upstream1* (than the further upstream sequence segments) only when viewed as translated to W/S, but not when sequence is viewed as R/Y. If in fact the regulatory sequences have a high CG content, the question arises whether the observed change in sequence matches (for W/S translated sequence data) is due to an increase in random matches within *upstream1* because of this change in W/S composition. In other words, does this change across the upstream reflect a meaningful change in frequency of these TFBS motifs across the positional segments, or is it simply due to random sequence matches due to compositional changes?

These two factors are of course not necessarily mutually exclusive. It may be that the increased density of regulatory sequences in the *upstream1* region is in actuality the reason for its markedly increased C/G composition. The following experiment takes into consideration the random expectation of sequence matches, and it will therefore become evident whether the observed increase in matches in the *upstream1* is due solely nucleotide compositional changes across the 5' upstream.

5.4.2 Binding motif representation

The results beg two important questions; How can the difference between the W/S- and R/Y-translated sequence representation profiles for the TFBS motifs be understood in terms of biological function? Also, what do the representation values and their relative changes across the upstream positional segments actually mean?

Regulatory protein binding to DNA depends on the presence of a specific binding site with the correct sequence and correct location and context within the genomic DNA. Protein-DNA recognition involves the two-step process of docking and probing. As previously described, indirect readout is thought to be the major contributor to protein-to-DNA binding and this relates to the docking phase.

The results for regulatory sequence representation may be considered in terms of the different possible outcomes; i.e. enhanced motifs, suppressed motifs or motifs present at the random level. Since regulatory elements function via the recognition and binding of proteins, involving docking and probing, the interpretation of these results should be in terms of these factors. In addition to this, for regulatory elements, location and context are also important and so any changes in representation can also be interpreted with these issues in mind. Therefore, it is useful to compare the results across the upstream sliding windows and genome-wide.

When a TFBS is suppressed, the implication is that there is an avoidance of its presence for a biological reason. In the case of transcription factor binding sites, the avoidance of this motif in the DNA (either in general or within certain locations) is likely related to the prevention of inappropriate protein binding.

When the TFBS motifs are enhanced within a particular location but not in other locations, this implies a positional preference of function. If a particular motif type is globally enhanced, this implies a wide-spread functional requirement. This may occur, for example, with

certain structural features. This is not expected to be the case for regulatory sequences though, since only certain genomic locations are responsible for gene regulation.

If on the other hand TFBS motifs occur at the random level (either generally or in specific locations), this implies by default that neither suppression nor enhancement are necessary. This suggests no biological requirement or issue. In the case of regulatory motifs random occurrence also implies that there would be no need to avoid inappropriate protein-DNA binding.

Bearing in mind these ideas it is possible to compare the different results and interpret them considering protein docking and probing and also positional context. Since the R/Y -translated TFBS motifs are 'globally' suppressed, there seems to be a general need to avoid them in DNA. This is probably in order to prevent random binding. The results for W/S sequences are very different. A close to random profile is seen everywhere except for *upstream1*, where the regulatory sequences are suppressed. This result implies that in every area tested except for *upstream1*, there is no need to avoid random binding, from the W/S perspective. Also, the results show that although within W/S -translated *upstream1* sequences, the TFBS motifs are suppressed, for R/Y this suppression is on average greater in *upstream1*.

Why are the R/Y- and W/S-translated sequence representation profiles so different? This question is of great importance, since the analysis of these datasets (derived from identical DNA sequences) has turned out very different results. The reason for the very different representation profiles is most probably due to differences in these properties of the bases and their effect and role within the regulatory regions and genomic DNA in general.

In chapter 3 and chapter 4 it was determined and discussed that the R/Y property of the base sequence most probably affects sequence structure more than does the W/S sequence, due to the R/Y influence on DNA flexibility verses rigidity. This is likely to affect the way in which the protein generally fits with the DNA sequence, thereby affecting the process of indirect readout and docking. The W/S property of the sequence may affect (to a greater extent than R/Y) direct readout, thereby influencing the process of probing. The results seen here also support this idea and seem to add strength to this argument.

The TFBS motifs are suppressed globally when the DNA is viewed as an R/Y sequence. For W/S sequences this is not the case. This apparent global suppression only for R/Y is possibly due to the role of the R/Y property of the bases in avoidance of inappropriate binding. It could be that R/Y motifs equivalent to regulatory elements must be suppressed in the DNA in general in order to prevent this. It is perhaps necessary to suppress these sequences 'everywhere' in order to generally avoid docking attempts by these proteins onto the DNA.

This brings us to the second phase of protein-DNA binding, that of probing. With the exception of the *upstream1* region, the TFBS motifs in W/S –translated sequences are present at the random level. Perhaps in these regions there is no need to avoid inappropriate binding of proteins at the direct readout level. Since docking comes first and also represents the majority of the protein-DNA interaction, the event of probing depends on the success of docking. Therefore there seems to be little need to suppress W/S motifs. Avoidance of the R/Y motifs (and protein docking) may just be enough.

The avoidance of inappropriate binding in genomic sequences in general can be described (according to these results) in terms of avoiding docking via the R/Y sequences. The W/S result which shows suppression of regulatory motifs only for *upstream1* requires further discussion.

Why for the W/S (translated sequence) results are the TFBS motifs suppressed in *upstream1*? In a place where a sequence is required for biological function, it would seem intuitively that it should be present above the randomly expected level. Yet here the opposite is true. The matches for W/S (translated) regulatory elements are twice as frequent in *upstream1* as the other positional segments. This suggests that their biological activity is predominantly in this region. Yet at the same time, in *upstream1* these same motifs are suppressed. Their suppression implies non-random presence.

It may be that within *upstream1*, further avoidance of incorrect binding is achieved by suppressing the presence of W/S motifs identical to regulatory elements. This would be an added measure beyond that of the general suppression of the R/Y motifs. This added measure may be in place since incorrect or inappropriate binding of regulatory protein (even weak or transient binding) to the DNA at this location may present a critical problem.

This further avoidance would be achieved by suppressing W/S regulatory motifs to avoid the probing step. In the further upstream sequence and the entire genome, only the avoidance of protein docking may be necessary. Therefore within *upstream1* where the activity of these regulatory sequences seems most important an extra and more specific measure may be required.

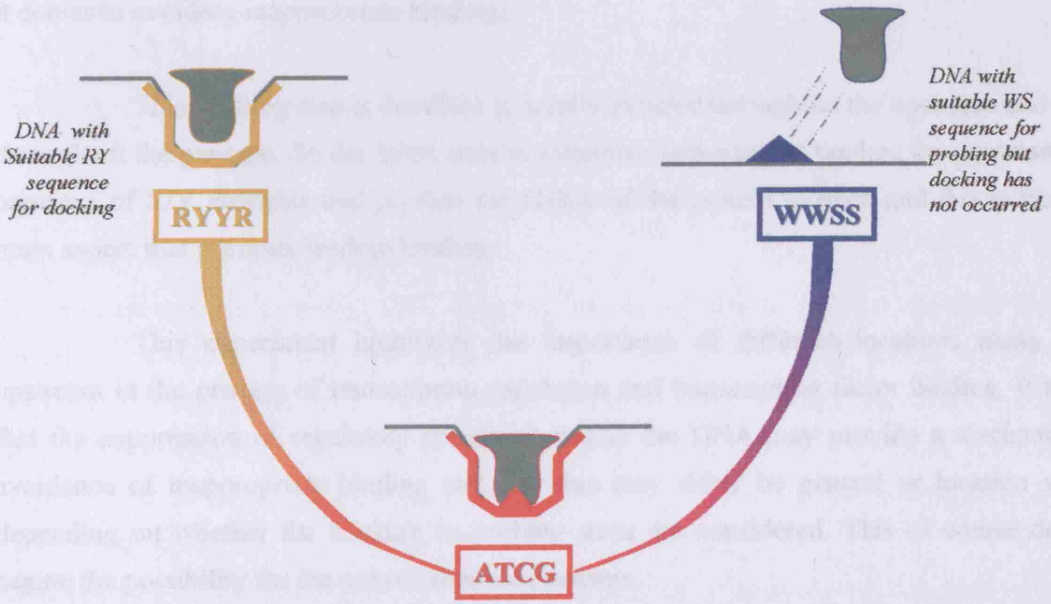
It is possible that in *upstream1*, the binding of regulatory proteins to the DNA can lead to very important changes in the level of transcript that is produced and that getting this right is crucial to the cell. *Upstream2*-to-*upstream5* may therefore be of a lesser importance for regulation. This could be the case for two related reasons; 1. *Upstream1* exists within the likely

promoter region and generally may have a much more prominent influence on transcription. 2. *Upstream1* may be generally more exposed to regulatory protein binding.

In conclusion, the relative differences seen in representation between R/Y- and W/S-translated sequences across the upstream are likely due to the different roles of these base properties play in the regulatory sequences with respect to the process of regulatory protein-DNA target binding.

It is now possible to conclude that the representation of TFBS motifs may actually be responsible (at least in part) for the process of avoidance of random binding of proteins to genomic DNA. Also, it seems that the representation of the regulatory sequences results in avoidance of inappropriate binding of regulatory proteins in specific regions where their biological function is likely to be most prevalent, namely *upstream1*.

The prevention of random binding may now be explained in terms of avoidance of protein docking (see figure 5.6). Generally docking may be prevented via suppression of certain R/Y motifs; namely those that resemble the regulatory elements. This suppression would result in an avoidance of the docking step.



Correct RY sequence for docking and correct WS sequence for probing; Results in successful protein-DNA binding

Figure 5.6:

This schematic diagram illustrates a possible model for regulatory protein-DNA docking and probing that depends on R/Y motifs and W/S motifs.

The arrangement of purine and pyrimidine bases in the DNA may determine the propensity for the DNA to allow protein docking. A suitable R/Y motif provides for the possibility of a protein to dock due to general fit interactions and the correct 'bend-ability' of the DNA.

A suitable W/S motif would allow for the second step of protein-DNA binding to take place i.e. probing. This is because a suitable W/S (together with the suitable R/Y) sequence promotes the more intricate binding interactions and in particular hydrogen bonding between amino acids and the nucleic acids. However, if docking has not taken place, a suitable W/S sequence for probing is ineffective.

Only if the correct R/Y sequence is present so that docking would occur will probing begin. Therefore, it is a combination of the correct R/Y sequence with the correct overlapping W/S sequence that results in successful docking and probing respectively. This means that the suitable ATCG sequence gives rise to successful protein-DNA binding.

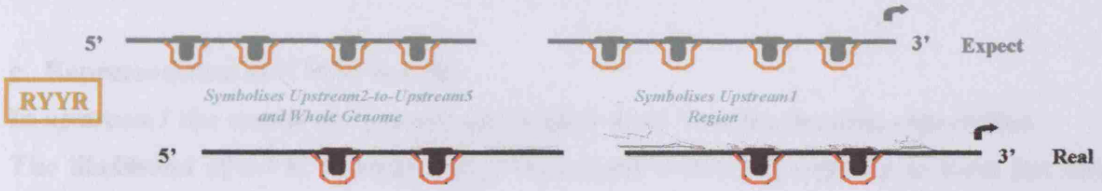
This means that the majority of the events of protein-DNA binding is eliminated since docking constitutes 2/3 of the bonding forces between protein and DNA. Also, critically docking is the first step in protein-DNA binding without which there cannot be a second step, i.e. probing. Most of the battle is thereby won (in any single protein-DNA binding event) when it comes to avoiding inappropriate binding.

This docking step is therefore generally avoided throughout the upstream and indeed throughout the genome. So the DNA tries to minimize non-specific binding by suppressing the presence of R/Y elements that provide the ability of the protein to dock and this is likely the main aspect that prevents random binding.

This experiment highlights the importance of different locations along the 5' upstream in the process of transcription regulation and transcription factor binding. It may be that the suppression of regulatory sequences within the DNA may provide a mechanism for avoidance of inappropriate binding and that this may either be general or location specific depending on whether the docking or probing steps are considered. This of course does not negate the possibility for the use of other mechanisms.

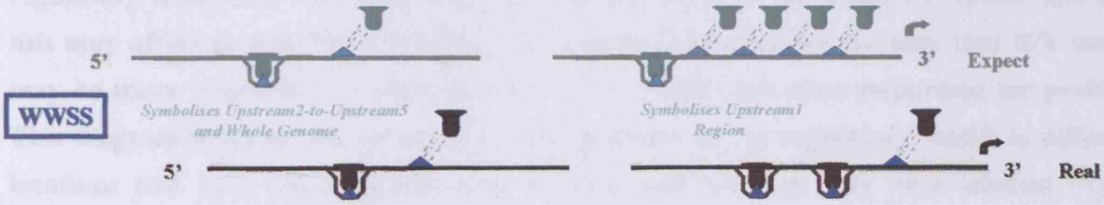
5.4.3 A model for avoidance of inappropriate transcription factor binding

So far, it has been discussed how the R/Y and W/S properties of the sequence may affect the binding of regulatory proteins to the DNA and how the process of inappropriate binding may be avoided. Now, these results will be consolidated with the ATCG results to produce a final model for this experiment (see figure 5.7).



a. Representation of R/Y motifs:

The first step is protein docking. According to the idea that R/Y motifs are more responsible for docking, it is seen that the real occurrence of these motifs is lower than expected in the whole of the upstream and in the whole genome. This could in theory be due to a need to prevent general large-scale docking events.

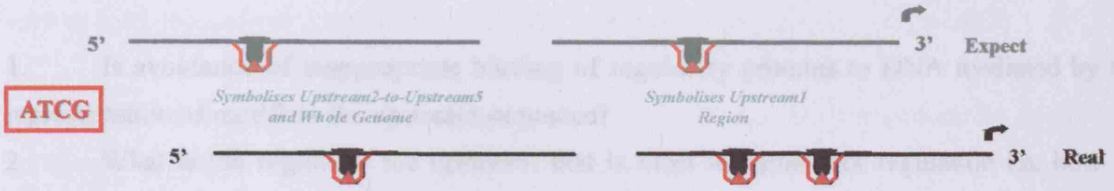


b. Representation of W/S motifs:

The second step is probing. Once docking has taken place, it is possible for the more intricate process of probing to occur. For this the correct W/S motifs may be more important. Some of the W/S sequence motifs overlap with the R/Y sequence that allowed for docking whilst others do not. In the case where the R/Y motif is present, the protein in theory could dock (as shown). Where this R/Y sequence is absent docking would not be possible, as shown.

In *upstream1* the regulatory motifs are present at a lower level than is randomly expected, whereas in the *upstream2*-to-*upstream5* and in the *whole genome* they are present at the random level. The random occurrence of W/S motifs in *upstream2*-to-*upstream5* and in the *whole genome* may be due to the fact that there is no need to suppress them since docking has already been suppressed via the R/Y motifs. Without docking there can be no probing.

In *upstream1* further suppression may be necessary in order to avoid the chance event of a correct W/S motif overlapping with a R/Y motif forming an unneeded regulatory element. In this region such a situation would pose a critical problem.



c. Representation of ATCG motifs:

In *upstream1* the motifs are present at a higher level than the random expectation. The likelihood of an R/Y motif and a W/S motif occurring together to form the correct ATCG sequence for a regulatory sequence is lower than their actual occurrence in *upstream1*. In *upstream2-to-upstream5* and the *whole genome* they occur at the random level.

Figure 5.7: A highly schematic diagram depicting a model for the avoidance of inappropriate protein-DNA binding and promotion of correct binding in the 5' upstream region of human genes. This model is based in the results of regulatory sequence representation in the upstream and genome-wide. This model describes that the regulatory motifs are relatively suppressed or enhanced in the different regions and how this may affect protein-DNA binding. The model is founded on the idea that R/Y motifs may be more important for docking whilst W/S motifs are more important for probing. This diagram helps to see the relative representation of the regulatory motifs in different locations and how this may relate to docking and probing. The area labelled expect illustrates the expected random occurrence of the regulatory motifs, whereas the area labelled real depicts their actual occurrence.

5.4.4 The overall message and questions that arise

In the introduction and aims of this project questions were asked regarding the upstream regulatory sequences, so the question now is have they been answered and if so to what extent? The initial questions were as follows:

1. Is avoidance of inappropriate binding of regulatory proteins to DNA mediated by the representation of motifs in the upstream sequence?
2. What is the region of the upstream that is most important for regulation i.e. how far upstream and is there any indication of a boundary?

These questions appear simple, but the results proved to be complicated. The initial (ATCG) analysis of regulatory motif representation suggested that genome-wide, random (or inappropriate) binding to motifs that are equivalent to regulatory sequences does not pose a problem at all. This conclusion was drawn because the motifs are present at the random expected level. In other words, there is no apparent attempt to avoid or suppress these motifs. This is not at all surprising since these motifs are not expected to possess any function as real regulatory elements in genomic DNA. However, the analysis of the upstream positional segments revealed an interesting insight into gene regulation.

This conclusion is based on the assumption that inappropriate binding would be a problem or a damaging event in these upstream regions. The enhancement of the regulatory sequence motifs in *upstream1* simply implied that this is where their biological activity is most prevalent, the non-random value implying biological relevance.

The interpretation of biological meaning for representation of motifs is of the essence here. So what is the real meaning of regulatory sequence representation? The answer to this lies in the theories behind the interpretation of under- /over- /random-representation of motifs in the DNA sequence. The above-given possibilities for representation and the linked biological requirement of the sequences are based on certain assumptions:

1. The first assumption is that sequences required biologically are enhanced. Although this seems intuitively to be true, it is not necessarily the case. In addition to this whilst enhancement may correlate with biological activity this activity may not necessarily relate directly to regulation. However, it is reasonable to assume that observations made for a set of regulatory elements do relate to the biological function of gene regulation.

2. Motifs that are suppressed (in the context of regulatory elements on the DNA within the upstream) are suppressed in order to avoid random binding by proteins.

3. That protein docking relates more to the R/Y sequence and probing relates to the W/S sequence.

The conclusions drawn for these experiments were therefore based on the interpretations that were rooted in the above-given assumptions. Although the general assumptions themselves seem reasonable, their truth is difficult to verify in practise.

Further insight was then obtained by looking at the sequences from two perspectives, namely, the R/Y and W/S properties of the bases. In chapter4 it was seen that the R/Y -translated sequence is more distant from randomness than the equivalent W/S -translated sequence with respect to its dinucleotide content, within the upstream and genomic DNA. This is likely due to the R/Y sequence being a more decisive proponent of DNA structure as previously discussed. The structural influence was proposed to become greater towards the TSS where regulatory sequences are more prevalent. This idea was put forward since the regulatory sequence structure is very important if that specific DNA sequence element will interact and permit protein docking. Also, indirect readout is thought to constitute the majority of the DNA-protein interaction.

It is important to note that for the regulatory element to operate effectively its sequence must be correct with respect to its R/Y bases, and also its W/S bases. Therefore if a particular sequence on the genomic DNA possesses the correct or identical R/Y sequence to a particular regulatory element, this does not mean that it will operate as a regulatory element (even in the slightest sense). This experiment is simply an attempt to understand the role of these two different properties of the bases within the regulatory elements and how this changes across the upstream.

Therefore when for example, frequency and representation of a W/S sequence (equivalent to real elements) within the genomic sequence is analysed, it is off course present many more times than the real ATCG sequence. This analysis though helps to see where this W/S sequence is avoided or enhanced, which then permits for deductions about the biology.

The recognition code for protein-DNA specificity remains elusive. It is known that for the regulatory proteins the majority of interactions occur within the major groove. Potential recognition patterns illustrate how the possibility for bonding is indeed greater in the major groove. Although recognition patterns reveal information about potential bonds formed within the major and minor groove, it remains unknown how docking and probing specifically occur with respect to the DNA sequence.

Other mechanisms for avoidance of inappropriate binding:

Regarding the issue of avoidance of random regulatory protein binding, there are likely to be mechanisms in place other than under-representation of regulatory motifs. Chromatin appears to play a role in transcription since transcriptionally active chromatin possesses an open structure. This probably gives the machinery access to active genes.

The closed chromatin structure may therefore in contrast prevent the access of this machinery including the upstream transcription factors. This may be regarded as a type of physical masking of the DNA. Therefore it would seem that chromosomal proteins may play a part in the transcriptional activity of genes. However, the widespread abundance of histones makes it unlikely that they are directly responsible for this specific role. Whilst they may be involved, it is likely that this is secondary, i.e. the consequence of a more specific event.

Despite the suppression of regulatory element motifs there may still be sequences that are identical to the regulatory elements that either occur randomly or that form a component of other elements, for example structural motifs. The transcription machinery and its specificity is of great importance and the mechanisms guarding it are likely complex.

One possible safeguard would be the existence of composite elements for gene regulation (Kel et al, 1995, Donaldson et al, 2007). These are elements that act together and are up to eighty bases apart. Therefore an element would require the existence of an additional signal in order to function. This would greatly reduce the likelihood of inappropriate binding, thereby strengthening specificity.

5.4.5 Limitations of the dataset and procedure

There is a question of defining the regulatory element and what constitutes a discrete binding domain. Since here the occurrence and distribution of regulatory motifs is tested and of particular interest are transcription factor binding regions of the DNA it is important to understand what the regulatory element actually contains. There are two issues;

1. Firstly, an element may not necessarily constitute a discrete binding domain to a protein. For example a larger element may contain some spacer sequence to which protein does not bind. I.e. a discrete binding domain may not have been isolated for the DNA element and therefore not all of the sequence necessarily constitutes a binding region.

2. The second issue is that of the nature of protein-DNA contacts. Within an element that constitutes a discrete protein binding domain (or region) there may be bases that come into contact with the protein whilst others do not. Furthermore some bases in the helix may be essential for the protein to bind whilst others may not be. Clearly protein-DNA contacts are highly complex.

Although the motifs used here ranged between five to ten bases in length, and this reduced the likelihood of spacer sequence being present, this problem may not have been eliminated. The second issue of necessary versus unnecessary bases presents an unknown problem. There are though nucleotides within or surrounding a regulatory element that do not interact directly with amino acids from the regulatory protein but do influence the activity of the element and its binding to the regulatory protein (Koudelka et al, 1987).

This dataset of regulatory motifs is not ideal. The perfect dataset would be a set of elements whereby the significance of each base in the element would be known. In other words, the specific interaction between the element and amino acids of the regulatory sequence would be defined, including its boundaries. It would be useful to know the significance of each base with respect to mutations within the regulatory element. This would also allow for motif searches to be carried out with mismatches, something that could not be done here.

It would have been preferable to have a larger dataset of regulatory motifs. Although there were 154 motifs, these varied from five to ten bases in length. It would have been better to have a larger dataset over a more extensive range of motif lengths and to carry out statistical analysis on separate groups according to length of sequence. The W/S and R/Y datasets of motifs were ten to twenty bases in length. Therefore whilst these could be compared to each other regarding their distribution within the upstream they could not be compared directly to the ATCG motifs which were shorter.

6. Summary, Perspectives & Further Work

In this chapter a comparison and consolidation of the different results and conclusions will be made in order to see how they may relate to the bigger picture. Potential further work will be outlined in response to questions that have been raised. Following each of the results chapters there was potential to investigate further along different paths of inquiry. Certain questions were followed up in this project whilst others were not. Some of those other possible trajectories of further work will also be described. The main emphasis though will be along the specific path of this project and its four major results sections.

6.1 Conclusions Summarised

The analysis of dinucleotide composition and representation showed that there were changes in these properties across the 5' upstream sequence of the human gene. Dinucleotide representation changes could be grouped into three coherent categories. 1. Dinucleotides that are suppressed in the upstream and become more suppressed towards the TSS. These all were found to contain one purine and one pyrimidine. 2. CpG is the only dinucleotide that becomes less suppressed towards the TSS. 3. Dinucleotides that are over-represented and become more enhanced towards the TSS. These all contain either two purines or two pyrimidines.

The upstream sequence was found to possess non-random characteristics with respect to dinucleotides. The trends in dinucleotide composition and representation across the 5' upstream sequence could be clearly categorized into purine-pyrimidine changes and into weak-strong changes. These trends and their groupings have an effect on DNA structure and suggest overall alterations in this structure towards the TSS.

Work carried out by El Hassan and Calladine, 1996 and Hunter 1993 shows the roll and slide angle ranges of DNA tracts and the resultant tendency of certain base steps to adopt A- / B-form DNA. This work outlined the tendency of DNA steps to produce flexible, rigid or bistable structures in accordance with roll and slide angle tendencies. Utilizing this (El Hassan and Calladine's experiments) together with the results of this project, it is concluded that there is a general increase in DNA rigidity and bistability and a simultaneous decrease in DNA flexibility towards the TSS.

Three types of dinucleotide were characterized from El Hassan and Calladine's experiments, namely; 1. Rigid dinucleotides, 2. Flexible dinucleotides and 3. Bistable dinucleotides. The rigid dinucleotides are those with narrow-range roll and slide angles. The flexible dinucleotides generally comprising of a pyrimidine and a purine base possess a wider range of roll and slide angles. These produce flexible tracts in that their potential conformations are between the A-form and B-form with multiple intermediates. In contrast, the bistable dinucleotides are able to adopt either high slide or low slide conformations but without the intermediates. They are therefore neither rigid (narrow-ranging) nor flexible (wide ranging) in there conformations, rather they are able to adopt two possible 'extreme' structures. These bistable dinucleotides are composed of two strong bases.

In genomic DNA there is a general tendency for flexible dinucleotides to be suppressed. In the upstream sequence this suppression is increased towards the TSS. In contrast stiff dinucleotide steps are seen to become increasingly enhanced towards the TSS. Bistable (SpS) steps are greatly increased in proportion and are increasingly enhanced towards the TSS.

This possibility to categorise changes in dinucleotide representation across the 10Kb upstream in a coherent manner according to purine-pyrimidine and weak-strong changes led to a separation of these properties for all subsequent experiments. Therefore the DNA sequence was translated from an ATCG sequence into two different types of sequence; namely 1. Purine and pyrimidine (R/Y) and 2. Weak and strong (W/S). Analyses of sequence were subsequently carried out on these two translations separately. This was carried out with the following question in mind; How does the cell 'read' and 'recognise' the upstream sequence and more specifically does the cell see the upstream sequence as a purine/pyrimidine sequence or as a weak/strong sequence in certain contexts?

The analyses that followed showed that the overall non-random characteristics of the upstream sequence changed in the TSS direction. This change was dependent on whether the sequence was viewed as a purine/pyrimidine (R/Y) sequence or as a weak/strong (W/S) sequence. It was found that the R/Y (translated) upstream sequence is more distant from randomness than the W/S (translated) sequence, throughout the 10Kb upstream. In addition, the R/Y sequence becomes more distant from the random model towards the TSS, whereas the W/S sequence becomes closer to the random model.

These results show a different level of significance for the R/Y and W/S sequence in the upstream. It is concluded that the R/Y sequence is more important in the 10Kb upstream region than the W/S sequence and that this importance increases towards the TSS. It has been suggested that the R/Y upstream sequence is less random than the equivalent W/S sequence due

to its role in determining the relative flexibility/stiffness of DNA. This in turn affects the process of protein docking to DNA. In the upstream sequence this structural role is especially pertinent.

The R/Y sequence has a greater influence over helical structure than the W/S sequence. Indirect readout constitutes the majority of the protein-DNA interaction and relies very much on helical structure. The determinants of DNA structure may therefore become even more important towards the TSS in the location where regulatory sequences are especially dense. The R/Y sequence therefore likely affects the ability of proteins to dock onto the DNA and affects the process of indirect readout to a greater extent than the W/S sequence.

It is also concluded that the W/S sequence is likely to be a generally greater determinant of direct readout and the R/Y sequence a greater determinant of indirect readout. This would explain the observation that the R/Y sequence is generally less random in the upstream than the W/S sequence. It also explains why the R/Y sequence becomes even more distant from randomness towards the TSS whereas the W/S becomes closer to randomness.

From the flexible, rigid and bistable dinucleotide properties it was further deduced that the combination of these is likely to play a key role in the recognition of regulatory proteins and their binding to the DNA. Flexible DNA has a higher tendency to bind proteins (e.g. nucleosomes). This is probably due to the greater ease with which proteins are able to initially dock to the DNA. Flexible dinucleotides (or base steps) were generally suppressed in the genomic DNA. Towards the TSS these were found to be even more suppressed indicating a general 'prevention' of protein docking in this region, to which transcription factors bind.

The other major difference between the intergenic DNA and the region close to the TSS was the great increase and enhancement (beyond randomness) of the bistable dinucleotides. Therefore it could be concluded that bistability is an important property of the DNA in relation to regulatory protein interaction.

Since rigidity is enhanced and flexibility suppressed towards the TSS, in general there should be less propensity for protein binding; therefore how do the regulatory proteins bind to the DNA close to the TSS? It was concluded that bistability may play a major role in this process, since the bistable steps allow for a transition of the DNA from high to low slide and critically without adopting the intermediate structures that are possible with the flexible dinucleotides. This bistable property may allow for a 'compensation' of the decreased flexibility of the DNA close to the TSS, making it possible for regulatory proteins to bind. Moreover, it is

likely that the bistability of this regulatory region DNA may be required specifically for regulatory protein binding and may even select specifically for these proteins.

The following general rules have been derived from the dinucleotide composition, representation and distance from randomness experiments of this project in conjunction with El Hassan's and Calladine's experiments described above;

- The R/Y sequence (of dinucleotides) has a greater influence over helical structure than the W/S sequence. Therefore the R/Y translated sequence likely affects the ability of proteins to dock onto the DNA and the process of indirect readout to a greater extent than the W/S sequence.
- It has also been proposed that the W/S sequence affects direct readout to a greater extent. This is supported by the H-bonding patterns within dinucleotide steps that are distinguished by their W/S arrangement within the minor groove and may also be partially true for the major groove.
- The SpS dinucleotide is exceptional since it generates bistability, i.e. this set of dinucleotides affect helical structure in a specific way (described above) regardless of their purine and pyrimidine content. This set of dinucleotides is very prevalent in regulatory regions.

In conclusion the W/S sequence is likely to be a greater determinant of direct readout and the R/Y sequence a greater determinant of indirect readout, with the exception of the SpS dinucleotides for which there are overlapping properties.

As a follow-up to the experiments described so far, sequence similarity across the upstream region was studied using a patterns analysis. Any changes in sequence similarity across the upstream towards the TSS would in theory indicate changes in level of functionality amongst the set of upstream sequences of human genes. Therefore it was expected that across the dataset of human genes the similarity of the sequence would increase towards the TSS.

Since the upstream sequence displayed different distance from randomness trends depending on whether it was viewed as the original (ATCG) sequence or translated into an R/Y and W/S sequence, this translation was also carried out for the sequence similarity experiments. It was discovered that as with distance from randomness trends across the upstream for the R/Y and W/S translations, sequence similarity trends were opposing. The R/Y sequence was found to become more similar towards the TSS across the set of upstream sequences. In contrast the W/S became less similar. This greatly strengthened the distance from randomness observation.

In addition, the sequence similarity result proved that the link between similarity and level of functionality is not simple.

Collectively therefore it could be concluded that the R/Y sequence is more important towards the TSS direction, since it becomes les random and more similar (or convergent) across the set of human genes. In contrast the W/S sequence becomes less important in this direction as it becomes more random and more divergent. Therefore it was reaffirmed that the W/S and R/Y translated sequence has different relative properties that are likely related to different (W/S and R/Y) functions across the upstream sequence towards the TSS. Since these changes towards the TSS occur in an opposing manner, this may be attributed to regulatory sequence therein.

So far it is evident that there are different and opposing trends in distance from randomness and sequence similarity across the upstream towards the TSS for the equivalent R/Y and W/S translated sequence. Since changes in these properties were observed towards the TSS and this could be attributed to the increased presence of regulatory sequence at this direction, the next step was an analysis of regulatory sequences. TFBS and their motifs were utilized.

It has already been stated that the R/Y sequence is likely to affect the process of protein docking to DNA to a greater extent and that the W/S sequence likely to affect protein probing to a greater extent. The TFBS motifs are DNA sequences to which regulatory proteins bind, i.e. these are locations where protein docking and probing occur.

The distribution and representation of these TFBS motifs in the upstream region of the gene is of particular interest since this would help reveal information on random or inappropriate binding of proteins to the DNA and how this is avoided. The essential and added dimension is the study of TFBS distribution and representation following translation of the sequence into R/Y and W/S. This is because it permits a separation of the analysis into the docking and probing steps of regulatory protein-DNA interaction.

The results revealed that for the R/Y (translated) upstream sequence the TFBS motifs were present at a much suppressed level throughout the upstream sequence. In contrast, for the W/S (translated) upstream sequence, the motifs were present at the random level with the exception of the sequence closest to the TSS where the motifs were suppressed.

It was concluded from this that avoidance of random binding of regulatory proteins in the upstream, likely occurs via the avoidance of the docking step. This is because the R/Y sequence equivalent to the TFBS motifs is generally avoided and the R/Y sequence is that

which primarily affects DNA flexure/stiffness. If docking, the initial step in protein-DNA interaction fails to occur; protein binding to DNA is altogether avoided. Since the W/S sequence equivalent to the TFBS motif is only suppressed close to the TSS and occurs at the random level elsewhere, probing the second step in protein-DNA binding is avoided only closest to the TSS.

These experiments in their totality revealed much information about the upstream sequence of the human gene. Sequence analysis and changes across the 10Kb upstream were utilised to draw conclusions about structure and function. This led to an understanding that the equivalent R/Y and W/S sequences (inherent within the ATCG sequence) possess a different level of importance to each other in the upstream region, especially towards the TSS. Furthermore it was possible to draw conclusions about the nature of transcription regulatory regions and their motifs. Most importantly the R/Y and W/S division of analysis has led to insight into how the regulatory DNA sequence may be 'read' and recognised by proteins with respect to docking and probing.

6.2 Outline of key issues & further work

Key issues that emerge from this project include;

- Changes in DNA sequence and structure across different positional regions of the upstream.
- The relative use or function of the R/Y and W/S nucleotides within the upstream.
- The distribution of regulatory elements within the upstream.
- The method of avoidance of inappropriate binding by regulatory proteins to the DNA and the process of specificity of regulatory protein binding to regulatory sequences.
- The possible relationship between indirect and direct readout during protein-DNA binding and the R/Y and W/S sequence arrangement in regulatory elements.

These issues are inter-related and in general terms the upstream region of human genes and their regulatory regions were of interest. Therefore the description of further work will be in response to these and will be presented as possible ways of addressing these issues.

6.2.1 Analysing sequence composition of the promoter and regulatory elements

The essential next experiment would be a more detailed sequence analysis of the promoter and regulatory elements. This would be in a similar vein to the work in this project, including a mononucleotide and dinucleotide analysis. This time though the promoter would be divided into regulatory binding site sequences, spacer sequences and also boundary regions. This would also provide an answer to the question of whether changes in sequence properties across the upstream in the downstream direction are due to the regulatory sequences or spacer sequences within the promoter.

The results of the dinucleotide composition and representation across the different positional segments of the upstream suggested that there are structural changes. It was concluded that there is a general reduced flexibility and increased bistability towards the start site region of the upstream sequence. Is there a difference between the binding site and spacer sequences in this respect?

Within the promotor there are regions that possess different functions. The promotor region approximately (on average) 300 to 50bp upstream of the TSS is thought to be the core region (Cooper et al, 2006). The region around 1000 to 500bp upstream of the start site is thought in many cases to contain negative elements. A division of the promotor into such regions may also be carried out and their analysis together with binding site and spacer sequence could provide clues as to structural differences.

Comparisons of human and mouse have revealed homologous blocks within promoters (Suzuki et al, 2004). The G/C content inside and outside of these blocks was found to be different. These types of analysis are a beginning to understanding promotor structure through the sequence.

Regulatory elements and specific binding site regions may be studied for their dinucleotide content. For example, it would be of interest to determine whether there is a tendency for certain structural features (such as a tendency for stiff and bistable dinucleotide steps) and whether these features are different to the sequences immediately surrounding the binding sites. If this is the case the structural features of the binding sites may be important in their targeting by regulatory proteins. Also, structural features are helpful for understanding protein-DNA interactions. All of the analyses described so far may also be repeated for enhancer elements.

6.2.2 More extensive analysis of regulatory motif distribution and representation within the upstream sequence

In chapter 5 the distribution and representation of regulatory motif matches across different positional segments of the upstream was studied. An extension of this would be to test the distribution of matches in conjunction with mismatches in the elements. These can be divided into those mismatches (or substitutions) that are known to cause loss (or reduction) of function and those that do not cause loss of function. The results could then be compared.

This experiment would be useful in producing more refined results because essential nucleotides within the regulatory element could be considered and differentiated from the non-essential. The results would be more complex to analyse, but nevertheless produce important results.

The regulatory motif distribution and representation experiment in the upstream sequence were carried out with promoter-derived regulatory motifs. A similar study may be repeated for enhancer-derived motifs in order to attempt to identify their possible distribution and representation in the upstream sequence. However a test sequence space far beyond the 10Kb upstream would be required for this.

6.2.3 Docking and Probing: R/Y verses W/S regulatory sequences

It has been proposed that the R/Y sequence has a greater effect on indirect readout whilst the W/S sequence affects direct readout to a greater extent. In this section will be discussed ways to distinguish between these base properties within the target DNA on direct and indirect readout. This type of study would determine the validity of the above hypothesis.

Analysis of protein-DNA binding may include several dimensions. The first is the properties of the DNA target motif. The second is the amino acid binding site. The third is the specific amino-acid base interaction.

Studies of protein-DNA interactions have focused on specific bonds between the individual amino acids and the bases. For example, H-bonds and Van der Waals bonds formed between an individual amino acids and a particular nucleotide have been ascertained.

Furthermore, the effect of amino acid mutations on protein-DNA interactions have been studied (Luscombe et al, 2002).

An analysis of DNA regulatory sequence mutations and the effects of these mutations on direct and indirect readout interactions may prove to be of great use in understanding these protein-DNA interactions. In order to distinguish direct and indirect readout, Van der Waals interaction and H-bonds may be considered, both the specific and non-specific binding to DNA. The amino acid interactions with specific bases would be of interest and the particular amino acid involved would be of lesser interest for this experiment.

In order to test the hypothesis of the relative influences of the R/Y and W/S sequences over indirect and direct readout, transition and transversion mutations of the target DNA sequence would have to be carried out and their effect on these interactions would be tested. The following is an outline of the experimental steps that may be involved:

1. Any amino acid interactions with specific bases in the regulatory sequence; Both H-bonds and Van der Waals bonds formed should be ascertained (as with work by Luscombe et al, 2001). Bonds formed with the phosphate backbone would be determined and differentiated from the more specific bonds formed with the base edges.
2. Nucleotide mutations would be carried out, ideally all possible mutations of target regulatory sequences, and the H-bonds and Van der Waals bonds formed should be ascertained.
3. The results can then be analysed to see the relative effects of transition and transversion mutations on H-bond formation and Van der Waals bonds, comparing specific and non-specific interactions. This may be done so that the net effect of transitions on these bonds would be compared to the net effect of transversions.
4. Following this the same results may be analysed but this time taking into consideration nearest neighbour effects.

The problem is that once docking is diminished probing is also diminished. This means that in theory if indirect readout is disrupted via target sequence mutation, the more specific process of direct readout may not take place. Therefore the following rules may be applied:

1. A target DNA sequence transition substitution should cause little or no change in protein docking, but should affect and reduce probing interactions:

2. A target DNA sequence transversion substitution should affect and disrupt both docking and probing.

This type of testing is far from perfect since the proof is obtained via indirect means, i.e. via a process of elimination that is incomplete. Another problem is that perhaps not all the bases in the target sequence are of (equal) importance for binding. It may even be that some bases (depending on their location) affect indirect readout more whilst others affect direct readout.

The last and most important issue to consider is that whilst protein-DNA interactions are subdivided into direct and indirect readouts, it may be that a diminishing of one type of interaction may undermine the other. Therefore hypothetically if direct readout interactions are disrupted, it may that the entire complex would be destabilised and it would be impossible to accurately determine indirect readout interactions. The occurrence of this in practice though may be avoided at least in part for the purpose of this experiment because only one base would be substituted at a time. This would probably minimise any destabilising effect.

This outlined methodology, via a process of elimination would reveal the relative importance of the individual nucleotides (and dinucleotides) and their R/Y and W/S properties on the docking and probing phases of protein-DNA binding. Therefore a large scale study of the binding patterns between regulatory proteins and regulatory elements combined with mutations of the bases would bring forth essential information about how these interactions occur.

6.2.4 Changes in other sequence property across the upstream

In this project changes in sequence composition, sequence similarity and regulatory motif distribution were analysed across the different positional regions of the upstream sequence, from the 5' to 3' end, up to the TSS. Other sequence properties can also be analysed in this way in order to better understand this region of the human genome.

For example, it is possible to ascertain the occurrence of different types of repeats in the upstream region of genes. Both direct repeats and other repeats such as LINES and Alus may be examined. This may be done to see if there is any particular order to repeat arrangement across the upstream sequence. The results may help to explain the possible roles of different repeat sequences in the 5' upstream region of genes. The structural aspects of these different repeat types may be studied utilising the dinucleotide composition and representation. Another example would be a similar an analysis of MARS and SARS.

6.2.5 Comparing the 5' upstream sequence of human and mouse and other eukaryotes

In order to gain a better understanding of the 5' upstream region of genes, human and mouse sequences may be compared. Do mouse upstream regions show similar trends across the upstream as were seen in this project for the human sequences? Although there is less data for the mouse it may be possible to study its sequences. General comparisons may be made, for example, differences in overall mononucleotides and dinucleotides between the mouse and the human.

Non-coding sequences (even regulatory) are known to be less highly conserved than coding sequences between the mouse and the human (Brickner et al, 1999). It may be possible to carry out a comparison of the 5' upstream region of a set human and mouse homologous genes. A nucleotide comparison and also a larger motif analysis can be applied. It would also be relevant to carry out such studies on primate genome and other eukaryotic genomes.

6.3 Overall Discussion

6.3.1 Regulatory sequences and their recognition by regulatory proteins

The way in which the amino acids of regulatory proteins recognize and read the specific nucleotides in DNA sequences to which they bind remains elusive. There are many possibilities for primary protein sequences folding to form a 3D protein, and there are many possible combinations of DNA sequence. Despite this the protein particle does in fact directly 'read out' the nucleotide sequence. Therefore there is likely to be some type of language (even if it varies between the protein classes) or rules via which the protein particle 'reads' in order to distinguish the motif to which it should bind over other sequence. This though is clearly complex.

There are different steps that can be taken in order to further the knowledge of how this process works. These methods include both practical and bioinformatic techniques. Perhaps taken together, these will yield a clearer picture. The way in which inappropriate regulatory protein to DNA is avoided is obviously a related topic. In order to better understand this it is necessary to also study the context in which regulatory sequences exist and operate in the genomic DNA.

In this project the upstream region was the subject of interest as a means of gaining a better understanding of transcription regulation. The DNA sequence was analysed so as to attain this objective. More specifically changes in the sequence properties at across different positions in the upstream region were looked at. The idea was to understand sequence composition which affects structure, sequence similarity within and across the different positional region, and the distribution and role of regulatory motifs therein.

6.3.2 Different layers of information within the DNA sequence

DNA is the molecule that holds the program for the building blocks of the cell and the complete organism. This is true since contained within it is the code for proteins which are the building blocks of the cell and also regulate its biochemical reactions. This information is stored within the genes in the form of the triplet code.

In addition to this are regions both upstream and downstream of the gene that determine its activity by regulating expression. This entire system is dependent on the recognition of the base sequence of the DNA molecule (in various ways, depending on whether coding, non-coding etc...) by other biomolecules. In other words the cell must be able to read the DNA appropriately. This is true not only for the coding regions but also the non-coding regions. Therefore the same molecule possesses different regions along its length, which the cell, via the biomolecules is able to distinguish and then read and interpret accurately in very different ways.

The chemical bases found at these different regions are the same but the 'language' or code so-to-speak must be different. Biomolecules therefore read and interpret the DNA and there must be some common factors regarding their recognition and binding to the DNA molecule, depending on location.

In this project the investigation of the R/Y and W/S base properties of the upstream sequence was a consequence of some interesting and unanticipated results from the dinucleotide analysis across different positional regions of the upstream. This theme was then continued throughout the project in order to investigate further the observed changes regarding R/Y and W/S bases.

The four bases of the DNA molecule possess some level of overlapping properties. There are two different types of sub-division of bases; a base may be either a purine or pyrimidine, which depends on its ring structure. The second property is that of its hydrogen

bonding capability. If the base is weak, it can only form two hydrogen bonds with its complementary base, whereas if it is strong it can form three hydrogen bonds. Also the weak and strong base-pairs produce different potential H-bonding patterns in the major and minor grooves of the helix. Therefore there are two different classes of property, each class contains two possibilities, and therefore there are four different combinations of chemical base that are generated from these two classes of property.

In a sense these can be regarded as two different sequences within a single sequence. It is the combined effect of these characteristics that makes each of the four bases unique and yet an individual base is not entirely unique because it shares each one of its two characteristics with another base.

This may seem simple, but perhaps this very simplicity and therefore these overlapping properties affect the way in which the DNA is 'read' by the cell machinery. This may be the key to the way in which DNA works. It is evident from this project that these properties play different roles in the sequence depending on location, context and function.

The process of protein binding to DNA is thought to occur in two steps; docking and probing. The first may involve some indirect readout and the second direct readout. It has been proposed that the R/Y and W/S sequence (at the dinucleotide level) affect differently the way in which protein-DNA binding occurs differently, R/Y having a greater influence on docking whilst W/S affecting probing.

6.3.3 Final remarks

These experiments have provided information about the R/Y and W/S arrangement within the sequence. It was seen that the different properties of the bases are not equally significant within and across the upstream sequence of the human gene and indeed within some other genomic regions.

The human genome possesses specific organisation and structure. Trends and changes in sequence properties across the 5' upstream sequence were observed in this work. These reflect structural and functional alterations. This project has also revealed important sequence features related to regulatory motifs to which proteins bind. The topic of regulatory protein-DNA binding is a subject of high level biological interest. Future work would comprise extending these experiments in order to understand better the 5' upstream region and regulatory sequences, their organisation and how they may function in terms of protein-DNA interactions.

Acknowledgements

Many thanks to Gad Yagil for his thorough review of this thesis and also for very generously giving of his time in discussion of this work. His insights and recommendations have been extremely helpful. Ian Tomlinsion has also read this thesis and provided some general comments.

I would like to thank Hadar Goldvag for valuable help with statistical analysis throughout the experiments. I would also like to thank Gavin Kelly for discussion with statistics in work leading up to this project.

I thank John Sgouros and Cancer Research UK for funding.

Appendix A

A.1 Mononucleotides: descriptive statistics

Summary charts provide the results of a descriptive statistics analysis for each of mononucleotide content (composition) in the 5' upstream region of human genes. There are 10 datasets; *upstream1*-to-*upstream10*, spanning 10Kb upstream of the start site of transcription (TSS), *upstream1* being closest to the TSS. Each dataset (*upstream1*-to-*upstream10*) contains 18,725, 1Kb DNA sequence fragments from sequence upstream of 18,725 different mRNA's.

Upstream1

	A	*T*	*C*	*G*
Mean	0.240	0.239	0.260	0.261
Standard Error	0.0004	0.0004	0.0004	0.0004
Median	0.240	0.237	0.257	0.256
Standard Deviation	0.0571	0.0587	0.0596	0.0589
Skewness	0.071	0.144	0.332	0.473
Count	18725	18725	18725	18725
Confidence Level(95.0%)	0.001	0.001	0.001	0.001

upstream2

	A	*T*	*C*	*G*
Mean	0.270	0.271	0.230	0.230
Standard Error	0.0004	0.0004	0.0004	0.0004
Median	0.272	0.272	0.225	0.225
Standard Deviation	0.0532	0.0527	0.0498	0.0493
Skewness	0.008	-0.037	0.668	0.762
Count	18725	18725	18725	18725
Confidence Level(95.0%)	0.001	0.001	0.001	0.001

upstream3

	A	*T*	*C*	*G*
Mean	0.274	0.275	0.226	0.225
Standard Error	0.0004	0.0004	0.0004	0.0003
Median	0.275	0.276	0.221	0.221
Standard Deviation	0.0537	0.0527	0.0480	0.0476
Skewness	0.062	0.065	0.714	0.634
Count	18725	18725	18725	18725
Confidence Level(95.0%)	0.001	0.001	0.001	0.001

upstream4

	A	*T*	*C*	*G*
Mean	0.275	0.276	0.225	0.224
Standard Error	0.0004	0.0004	0.0004	0.0004
Median	0.275	0.276	0.220	0.220
Standard Deviation	0.0542	0.0537	0.0481	0.0482
Skewness	0.111	0.117	0.773	0.662
Count	18725	18725	18725	18725
Confidence Level(95.0%)	0.001	0.001	0.001	0.001

upstream5

	A	*T*	*C*	*G*
Mean	0.276	0.276	0.224	0.224
Standard Error	0.0004	0.0004	0.0004	0.0004
Median	0.276	0.276	0.220	0.219
Standard Deviation	0.0550	0.0548	0.0482	0.0483
Skewness	0.136	0.120	0.723	0.750
Count	18725	18725	18725	18725
Confidence Level(95.0%)	0.001	0.001	0.001	0.001

upstream6

	A	*T*	*C*	*G*
Mean	0.276	0.276	0.224	0.224
Standard Error	0.0004	0.0004	0.0004	0.0004
Median	0.275	0.276	0.220	0.219
Standard Deviation	0.0563	0.0549	0.0482	0.0483
Skewness	0.210	0.136	0.788	0.726
Count	18725	18725	18725	18725
Confidence Level(95.0%)	0.001	0.001	0.001	0.001

upstream7

	A	*T*	*C*	*G*
Mean	0.276	0.276	0.225	0.224
Standard Error	0.0004	0.0004	0.0004	0.0004
Median	0.275	0.275	0.220	0.219
Standard Deviation	0.0565	0.0555	0.0488	0.0482
Skewness	0.230	0.170	0.782	0.662
Count	18725	18725	18725	18725
Confidence Level(95.0%)	0.001	0.001	0.001	0.001

upstream8

	A	*T*	*C*	*G*
Mean	0.277	0.276	0.223	0.224
Standard Error	0.0004	0.0004	0.0004	0.0004
Median	0.276	0.275	0.219	0.219
Standard Deviation	0.0572	0.0556	0.0482	0.0479
Skewness	0.239	0.208	0.724	0.671
Count	18725	18725	18725	18725
Confidence Level(95.0%)	0.001	0.001	0.001	0.001

upstream9

	A	*T*	*C*	*G*
Mean	0.276	0.276	0.224	0.224
Standard Error	0.0004	0.0004	0.0004	0.0004
Median	0.276	0.275	0.219	0.219
Standard Deviation	0.0569	0.0557	0.0481	0.0483
Skewness	0.214	0.206	0.704	0.750
Count	18725	18725	18725	18725
Confidence Level(95.0%)	0.001	0.001	0.001	0.001

upstream10

	A	*T*	*C*	*G*
Mean	0.276	0.276	0.223	0.224

Standard Error	0.0004	0.0004	0.0004	0.0004
Median	0.275	0.276	0.219	0.219
Standard Deviation	0.0567	0.0564	0.0480	0.0485
Skewness	0.236	0.194	0.644	0.756
Count	18725	18725	18725	18725
Confidence Level(95.0%)	0.001	0.001	0.001	0.001

A.2 Mononucleotides: Quantitative statistics

An ANOVA (single factor) analysis was carried out for the occurrence (or proportion) of a given mononucleotide within the ten upstream datasets; *upstream1*-to-*upstream10*. The ANOVA shows a comparison of mononucleotide content across the upstream segments.

Null hypothesis, Ho: the samples *(upstream1*-to-*upstream10)* are drawn from the same underlying probability distribution

Alternative Hypothesis, H_1: the underlying probability distributions for *upstream1*-to-*upstream10* are not the same for all samples. I.e. at least one of the populations has a mean not equal to the others.

The result reveals that the null hypothesis is rejected at the 5% level of significance for each of the four mononucleotides (see the summary table below). Therefore within the ten datasets, *upstream1*-to-*upstream10*, at least one of the datasets was significantly different to the others.

ANOVA summary table: single factor analysis at 5% level of significance			
For ten upstream datasets: upstream1-to-upstream10			
Mononucleotide	**Probability**	**Ho (reject/accept)**	**datasets – (inference)**
Adenine	0.000000	Reject	Different
Thymine	0.000000	Reject	Different
Cytosine	0.000000	Reject	Different
Guanine	0.000000	Reject	Different

The following data-tables show a more detailed breakdown of the ANOVA results given the summary table above for each of the individual four mononucleotides.

Adenine: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	614.55	9.000	68.284	23324	0.000	1.8799
Within Groups	548.17	187240	0.003			
Total	1162.7	187249				

Thymine: ANOVA Single Factor analysis

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	23.23	9.000	2.581	850	0.000	1.8799
Within Groups	568.33	187240	0.003			
Total	591.6	187249				

Cytosine: ANOVA Single Factor analysis

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	21.09	9.000	2.343	952	0.000	1.8799
Within Groups	460.68	187240	0.002			
Total	481.8	187249				

Guanine: ANOVA Single Factor analysis

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	23.07	9.000	2.564	1048	0.000	1.8799
Within Groups	457.92	187240	0.002			
Total	481.0	187249				

T-tests of the mononucleotide content of adjacent upstream positional segments were used to determine whether two samples are likely to have come from the same two underlying populations that have the same mean. In other words, *upstream1* was compared with *upstream2*, *upstream2* with *upstrream3* etc…It was not assumed that the populations contained an equal variance. These T-tests were two-tailed and carried out at the 5% significance level.

Null hypothesis, Ho: the samples from the two postionally adjacent upstream ($upstream_x$ and $upstream_{x+1}$) datasets are drawn from the same underlying probability distribution.

Alternative Hypothesis, H_1: the underlying probability distributions are not the same for these two adjacent sets of samples.

SUMMARY TABLE: Comparison of Adjacent upstream segments: two-tailed T-tests at 5% level. P values highlighted are for pairs of upstream datasets found to be significantly different									
Adjacent upstream datasets / **Dinucleotides**	upstream9/10	upstream8/9	upstream7/8	upstream6/7	upstream5/6	upstream4/5	upstream3/4	upstream2/3	upstream1/2
A	0.2439	0.2685	0.0437	0.8909	0.6938	0.0994	0.0424	0.0000	0.0000
T	0.8116	0.9205	0.8538	0.9319	0.6469	0.8987	0.1933	0.0000	0.0000
C	0.9333	0.5382	0.0232	0.8628	0.7386	0.3081	0.0000	0.0000	0.0000
G	0.2435	0.417	0.7583	0.9114	0.7897	0.4812	0.0000	0.0000	0.0000

The summary table above shows the results for these T-tests of adjacent segments across the 10Kb upstream. For all four mononucleotides the comparison between pairs of datasets revealed that;

- *upstream1* and *upstream2* are significantly different.
- *upstream2* and *upstream3* are significantly different.

Also, the descriptive statistics showed that there is an increased difference between the upstream segments towards the TSS. It therefore appears from these results that the three datasets derived from adjacent upstream regions (*upstream1*, *upstream2*, *upstream3*) are different to each other with respect to all four mononucleotides and that this is increasingly the case in the direction of the TSS.

The remaining seven sequence datasets (spanning a total of 7Kb), when analyzed in this adjacent location, pair-wise manner, in most cases showed that the datasets were the same. Hence the null hypothesis was accepted for these comparisons. Therefore for the most-part this 6Kb sequence possesses similar mononucleotide content. There is though one exception for two of the mononucleotides. For both adenine and cytosine the pair-wise comparison of the *upstream7* with the *upstream8* dataset revealed that these were significantly different.

The following data-tables show a more detailed breakdown of the T-test results (shown in the summary table above) for each of the individual four mononucleotides.

Adenine
Results for two-tailed T-tests of upstream adjacent datasets: 5% level of significance

	Probability	Ho (reject/accept)	datasets -inference
UPSTREAM9 / upstream10	0.2439	accept	same
UPSTREAM8 / upstream9	0.2685	accept	same
UPSTREAM7 / upstream8	0.0437	reject	different
UPSTREAM6 / upstream7	0.8909	accept	same
UPSTREAM5 / upstream6	0.6938	accept	same
UPSTREAM4 / upstream5	0.0994	accept	same
UPSTREAM3 / upstream4	0.0424	reject	different
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

Thymine

Results for two-tailed T-tests of adjacent upstream datasets:
5% level of significance

	Probability	Ho (reject/accept)	datasets - same/different)
UPSTREAM9 / upstream10	0.8116	accept	same
UPSTREAM8 / upstream9	0.9205	accept	same
UPSTREAM7 / upstream8	0.8538	accept	same
UPSTREAM6 / upstream7	0.9319	accept	same
UPSTREAM5 / upstream6	0.6469	accept	same
UPSTREAM4 / upstream5	0.8987	accept	same
UPSTREAM3 / upstream4	0.1933	accept	same
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

Cytosine

Results for two-tailed T-tests of adjacent upstream datasets
5% level of significance

	Probability	Ho (reject/accept)	datasets - same/different)
UPSTREAM9 / upstream10	0.9333	accept	same
UPSTREAM8 / upstream9	0.5382	accept	same
UPSTREAM7 / upstream8	0.0232	reject	different
UPSTREAM6 / upstream7	0.8628	accept	same
UPSTREAM5 / upstream6	0.7386	accept	same
UPSTREAM4 / upstream5	0.3081	accept	same
UPSTREAM3 / upstream4	0.0552	accept	same
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

Guanine

Results for two-tailed T-tests of adjacent upstream datasets
5% level of significance

	Probability	Ho (reject/accept)	datasets - same/different)
UPSTREAM9 / upstream10	0.2435	accept	same
UPSTREAM8 / upstream9	0.4170	accept	same
UPSTREAM7 / upstream8	0.7583	accept	same
UPSTREAM6 / upstream7	0.9114	accept	same
UPSTREAM5 / upstream6	0.7897	accept	same
UPSTREAM4 / upstream5	0.4812	accept	same
UPSTREAM3 / upstream4	0.0705	accept	same
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

A.3 Dinucleotide composition: Descriptive statistics

Summary charts provide the results of a descriptive statistics analysis for each of dinucleotide content (composition) in the 5' upstream region of human genes. There are 10 datasets; *upstream1*-to-*upstream10*, spanning 10Kb upstream of the start site of transcription (TSS), *upstream1* being closest to the TSS. Each dataset (*upstream1*-to-*upstream10*) contains 18,725, 1Kb DNA sequence fragments from sequence upstream of 18,725 different mRNA's.

Upstream1

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
Mean	0.071	0.049	0.048	0.073	0.042	0.070	0.060	0.067
Standard Error	0.0002	0.0002	0.0001	0.0001	0.0002	0.0002	0.0001	0.0001
Median	0.068	0.046	0.047	0.072	0.039	0.066	0.059	0.066
Standard Deviation	0.0317	0.0241	0.0106	0.0142	0.0224	0.0323	0.0130	0.0151
Skewness	0.487	0.719	1.233	0.820	0.844	0.617	0.844	0.877
Count	18725	18725	18725	18725	18725	18725	18725	18725
Confidence Level(95.0%)	0.0005	0.0003	0.0002	0.0002	0.0003	0.0005	0.0002	0.0002

Upstream1

	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
Mean	0.067	0.071	0.083	0.039	0.060	0.048	0.069	0.083
Standard Error	0.0001	0.0001	0.0003	0.0002	0.0001	0.0001	0.0002	0.0003
Median	0.067	0.071	0.078	0.032	0.059	0.048	0.066	0.078
Standard Deviation	0.0134	0.0141	0.0351	0.0304	0.0135	0.0110	0.0275	0.0356
Skewness	0.661	0.587	0.818	1.001	1.321	1.726	0.718	0.884
Count	18725	18725	18725	18725	18725	18725	18725	18725
Confidence Level(95.0%)	0.0002	0.0002	0.0005	0.0004	0.0002	0.0002	0.0004	0.0005

Upstream2

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
Mean	0.084	0.062	0.050	0.073	0.053	0.085	0.060	0.072
Standard Error	0.0002	0.0002	0.0001	0.0001	0.0002	0.0002	0.0001	0.0001
Median	0.083	0.061	0.050	0.072	0.052	0.083	0.060	0.072
Standard Deviation	0.0329	0.0231	0.0105	0.0152	0.0225	0.0329	0.0137	0.0135
Skewness	0.387	0.536	1.835	1.637	0.573	0.380	1.062	1.520
Count	18725	18725	18725	18725	18725	18725	18725	18725
Confidence Level(95.0%)	0.0005	0.0003	0.0002	0.0002	0.0003	0.0005	0.0002	0.0002

Upstream2

	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
Mean	0.072	0.073	0.066	0.018	0.060	0.050	0.053	0.066
Standard Error	0.0001	0.0001	0.0002	0.0001	0.0001	0.0001	0.0001	0.0002
Median	0.072	0.073	0.061	0.013	0.059	0.050	0.051	0.061
Standard Deviation	0.0126	0.0150	0.0285	0.0167	0.0144	0.0107	0.0192	0.0282
Skewness	1.314	0.559	1.326	2.685	2.561	2.853	1.094	1.320
Count	18725	18725	18725	18725	18725	18725	18725	18725
Confidence Level(95.0%)	0.0002	0.0002	0.0004	0.0002	0.0002	0.0002	0.0003	0.0004

Upstream3								
	ApA	**ApT**	**ApC**	**ApG**	**TpA**	**TpT**	**TpC**	**TpG**
Mean	0.086	0.065	0.050	0.073	0.055	0.086	0.060	0.074
Standard Error	0.0002	0.0002	0.0001	0.0001	0.0002	0.0002	0.0001	0.0001
Median	0.084	0.064	0.050	0.072	0.054	0.085	0.060	0.073
Standard Deviation	0.0340	0.0231	0.0106	0.0149	0.0226	0.0335	0.0140	0.0128
Skewness	0.428	0.522	2.386	0.486	0.610	0.428	1.297	0.854
Count	18725	18725	18725	18725	18725	18725	18725	18725
Confidence Level(95.0%)	0.0005	0.0003	0.0002	0.0002	0.0003	0.0005	0.0002	0.0002

Upstream3								
	CpA	**CpT**	**CpC**	**CpG**	**GpA**	**GpT**	**GpC**	**GpG**
Mean	0.073	0.074	0.064	0.015	0.060	0.051	0.051	0.064
Standard Error	0.0001	0.0001	0.0002	0.0001	0.0001	0.0001	0.0001	0.0002
Median	0.073	0.073	0.059	0.011	0.059	0.050	0.049	0.059
Standard Deviation	0.0129	0.0153	0.0272	0.0141	0.0137	0.0103	0.0181	0.0268
Skewness	1.781	0.756	1.310	3.229	0.836	1.535	1.068	1.239
Count	18725	18725	18725	18725	18725	18725	18725	18725
Confidence Level(95.0%)	0.0002	0.0002	0.0004	0.0002	0.0002	0.0001	0.0003	0.0004

Upstream4								
	ApA	**ApT**	**ApC**	**ApG**	**TpA**	**TpT**	**TpC**	**TpG**
Mean	0.086	0.065	0.050	0.073	0.055	0.086	0.060	0.074
Standard Error	0.0003	0.0002	0.0001	0.0001	0.0002	0.0003	0.0001	0.0001
Median	0.084	0.064	0.050	0.072	0.054	0.085	0.059	0.074
Standard Deviation	0.0342	0.0240	0.0107	0.0150	0.0235	0.0344	0.0138	0.0131
Skewness	0.451	0.912	1.568	0.606	0.981	0.478	0.890	1.224
Count	18725	18725	18725	18725	18725	18725	18725	18725
Confidence Level(95.0%)	0.0005	0.0003	0.0002	0.0002	0.0003	0.0005	0.0002	0.0002

Upstream4								
	CpA	**CpT**	**CpC**	**CpG**	**GpA**	**GpT**	**GpC**	**GpG**
Mean	0.074	0.073	0.063	0.014	0.060	0.051	0.051	0.063
Standard Error	0.0001	0.0001	0.0002	0.0001	0.0001	0.0001	0.0001	0.0002
Median	0.074	0.073	0.059	0.011	0.059	0.050	0.049	0.059
Standard Deviation	0.0128	0.0150	0.0274	0.0137	0.0137	0.0108	0.0180	0.0268
Skewness	0.935	0.466	1.596	3.115	1.177	1.997	1.003	1.232
Count	18725	18725	18725	18725	18725	18725	18725	18725
Confidence Level(95.0%)	0.0002	0.0002	0.0004	0.0002	0.0002	0.0002	0.0003	0.0004

Upstream5								
	ApA	**ApT**	**ApC**	**ApG**	**TpA**	**TpT**	**TpC**	**TpG**
Mean	0.087	0.066	0.050	0.073	0.055	0.087	0.060	0.074
Standard Error	0.0003	0.0002	0.0001	0.0001	0.0002	0.0003	0.0001	0.0001
Median	0.085	0.065	0.050	0.072	0.054	0.085	0.059	0.073
Standard Deviation	0.0350	0.0236	0.0107	0.0151	0.0232	0.0349	0.0141	0.0132
Skewness	0.474	0.684	1.675	0.543	0.898	0.475	0.869	1.256
Count	18725	18725	18725	18725	18725	18725	18725	18725
Confidence Level(95.0%)	0.0005	0.0003	0.0002	0.0002	0.0003	0.0005	0.0002	0.0002

Upstream5								
	CpA	**CpT**	**CpC**	**CpG**	**GpA**	**GpT**	**GpC**	**GpG**

Mean	0.074	0.073	0.063	0.014	0.060	0.050	0.051	0.063
Standard Error	0.0001	0.0001	0.0002	0.0001	0.0001	0.0001	0.0001	0.0002
Median	0.073	0.073	0.058	0.011	0.059	0.050	0.049	0.058
Standard Deviation	0.0130	0.0153	0.0272	0.0140	0.0139	0.0109	0.0183	0.0270
Skewness	0.873	0.453	1.389	3.333	0.994	1.774	1.068	1.357
Count	18725	18725	18725	18725	18725	18725	18725	18725
Confidence Level(95.0%)	0.0002	0.0002	0.0004	0.0002	0.0002	0.0002	0.0003	0.0004

Upstream6

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
Mean	0.087	0.066	0.051	0.073	0.055	0.086	0.060	0.074
Standard Error	0.0003	0.0002	0.0001	0.0001	0.0002	0.0003	0.0001	0.0001
Median	0.084	0.065	0.050	0.072	0.054	0.085	0.059	0.073
Standard Deviation	0.0357	0.0238	0.0111	0.0151	0.0233	0.0348	0.0142	0.0134
Skewness	0.554	0.607	2.051	0.497	0.876	0.507	0.903	1.495
Count	18725	18725	18725	18725	18725	18725	18725	18725
Confidence Level(95.0%)	0.0005	0.0003	0.0002	0.0002	0.0003	0.0005	0.0002	0.0002

Upstream6

	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
Mean	0.074	0.073	0.063	0.014	0.060	0.050	0.051	0.063
Standard Error	0.0001	0.0001	0.0002	0.0001	0.0001	0.0001	0.0001	0.0002
Median	0.074	0.072	0.058	0.011	0.059	0.050	0.048	0.058
Standard Deviation	0.0132	0.0154	0.0272	0.0138	0.0140	0.0111	0.0182	0.0272
Skewness	1.295	0.464	1.473	3.342	0.979	2.379	1.055	1.331
Count	18725	18725	18725	18725	18725	18725	18725	18725
Confidence Level(95.0%)	0.0002	0.0002	0.0004	0.0002	0.0002	0.0002	0.0003	0.0004

Upstream7

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
Mean	0.087	0.066	0.051	0.073	0.055	0.086	0.060	0.074
Standard Error	0.0003	0.0002	0.0001	0.0001	0.0002	0.0003	0.0001	0.0001
Median	0.084	0.065	0.050	0.072	0.054	0.084	0.059	0.074
Standard Deviation	0.0360	0.0240	0.0115	0.0151	0.0236	0.0354	0.0142	0.0130
Skewness	0.572	0.889	2.451	0.476	1.086	0.527	0.967	0.568
Count	18725	18725	18725	18725	18725	18725	18725	18725
Confidence Level(95.0%)	0.0005	0.0003	0.0002	0.0002	0.0003	0.0005	0.0002	0.0002

Upstream7

	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
Mean	0.074	0.073	0.063	0.014	0.060	0.050	0.051	0.063
Standard Error	0.0001	0.0001	0.0002	0.0001	0.0001	0.0001	0.0001	0.0002
Median	0.074	0.073	0.058	0.011	0.059	0.050	0.049	0.059
Standard Deviation	0.0136	0.0155	0.0277	0.0139	0.0139	0.0108	0.0183	0.0268
Skewness	1.597	0.463	1.534	3.448	0.945	1.117	1.030	1.265
Count	18725	18725	18725	18725	18725	18725	18725	18725
Confidence Level(95.0%)	0.0002	0.0002	0.0004	0.0002	0.0002	0.0002	0.0003	0.0004

Upstream8

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
Mean	0.087	0.066	0.051	0.073	0.055	0.086	0.060	0.074
Standard Error	0.0003	0.0002	0.0001	0.0001	0.0002	0.0003	0.0001	0.0001
Median	0.085	0.065	0.050	0.072	0.055	0.084	0.059	0.074

Standard Deviation	0.0365	0.0238	0.0111	0.0151	0.0233	0.0355	0.0139	0.0137
Skewness	0.562	0.632	1.395	0.568	0.780	0.539	0.712	1.379
Count	18725	18725	18725	18725	18725	18725	18725	18725
Confidence Level(95.0%)	0.0005	0.0003	0.0002	0.0002	0.0003	0.0005	0.0002	0.0002

Upstream8

	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
Mean	0.074	0.073	0.062	0.014	0.060	0.050	0.050	0.063
Standard Error	0.0001	0.0001	0.0002	0.0001	0.0001	0.0001	0.0001	0.0002
Median	0.074	0.072	0.058	0.011	0.059	0.050	0.048	0.058
Standard Deviation	0.0132	0.0153	0.0273	0.0136	0.0139	0.0114	0.0183	0.0266
Skewness	0.782	0.344	1.503	3.344	1.062	1.989	1.049	1.200
Count	18725	18725	18725	18725	18725	18725	18725	18725
Confidence Level(95.0%)	0.0002	0.0002	0.0004	0.0002	0.0002	0.0002	0.0003	0.0004

Upstream9

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
Mean	0.087	0.066	0.051	0.073	0.055	0.086	0.060	0.074
Standard Error	0.0003	0.0002	0.0001	0.0001	0.0002	0.0003	0.0001	0.0001
Median	0.085	0.065	0.050	0.072	0.054	0.084	0.059	0.074
Standard Deviation	0.0363	0.0237	0.0112	0.0151	0.0232	0.0356	0.0140	0.0136
Skewness	0.552	0.545	2.387	0.384	0.642	0.546	1.000	1.844
Count	18725	18725	18725	18725	18725	18725	18725	18725
Confidence Level(95.0%)	0.0005	0.0003	0.0002	0.0002	0.0003	0.0005	0.0002	0.0002

Upstream9

	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
Mean	0.074	0.073	0.062	0.014	0.060	0.051	0.051	0.063
Standard Error	0.0001	0.0001	0.0002	0.0001	0.0001	0.0001	0.0001	0.0002
Median	0.074	0.072	0.058	0.011	0.059	0.050	0.049	0.058
Standard Deviation	0.0134	0.0153	0.0269	0.0135	0.0140	0.0116	0.0183	0.0271
Skewness	1.375	0.533	1.269	3.273	0.862	2.876	1.046	1.358
Count	18725	18725	18725	18725	18725	18725	18725	18725
Confidence Level(95.0%)	0.0002	0.0002	0.0004	0.0002	0.0002	0.0002	0.0003	0.0004

Upstream10

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
Mean	0.087	0.066	0.050	0.073	0.055	0.087	0.060	0.074
Standard Error	0.0003	0.0002	0.0001	0.0001	0.0002	0.0003	0.0001	0.0001
Median	0.084	0.065	0.050	0.072	0.054	0.085	0.059	0.074
Standard Deviation	0.0361	0.0241	0.0111	0.0154	0.0235	0.0360	0.0139	0.0136
Skewness	0.547	0.809	1.702	0.885	1.074	0.548	0.876	1.392
Count	18725	18725	18725	18725	18725	18725	18725	18725
Confidence Level(95.0%)	0.0005	0.0003	0.0002	0.0002	0.0003	0.0005	0.0002	0.0002

Upstream10

	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
Mean	0.074	0.073	0.062	0.014	0.060	0.051	0.051	0.063
Standard Error	0.0001	0.0001	0.0002	0.0001	0.0001	0.0001	0.0001	0.0002
Median	0.073	0.072	0.058	0.011	0.059	0.050	0.048	0.058
Standard Deviation	0.0133	0.0153	0.0267	0.0138	0.0142	0.0113	0.0184	0.0271
Skewness	0.970	0.408	1.232	3.422	1.555	1.971	1.091	1.316
Count	18725	18725	18725	18725	18725	18725	18725	18725

Confidence Level(95.0%)	0.0002	0.0002	0.0004	0.0002	0.0002	0.0002	0.0003	0.0004

The following charts and their associated data-tables show the changes in dinucleotide proportions across the 10Kb upstream sequence (as separate datasets: *upstream1*-to-*upstream10*). The results are shown for each of the sixteen individual dinucleotides.

Dinucleotide proportions (median) for different upstream positional segments

Upstream Sequence \ Dinucleotides	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
Upstream10	0.084	0.065	0.050	0.072	0.054	0.085	0.059	0.074
Upstream9	0.085	0.065	0.050	0.072	0.054	0.084	0.059	0.074
Upstream8	0.085	0.065	0.050	0.072	0.055	0.084	0.059	0.074
Upstream7	0.084	0.065	0.050	0.072	0.054	0.084	0.059	0.074
Upstream6	0.084	0.065	0.050	0.072	0.054	0.085	0.059	0.073
Upstream5	0.085	0.065	0.050	0.072	0.054	0.085	0.059	0.073
Upstream4	0.084	0.064	0.050	0.072	0.054	0.085	0.059	0.074
Upstream3	0.084	0.064	0.050	0.072	0.054	0.085	0.060	0.073

Upstream2	0.083	0.061	0.050	0.072	0.052	0.083	0.060	0.072
Upstream1	0.068	0.046	0.047	0.072	0.039	0.066	0.059	0.066

Dinucleotide proportions (median) for different upstream positional segments

Upstream sequence \ Dinucleotides	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
Upstream10	0.073	0.072	0.058	0.011	0.059	0.050	0.048	0.058
Upstream9	0.074	0.072	0.058	0.011	0.059	0.050	0.049	0.058
Upstream8	0.074	0.072	0.058	0.011	0.059	0.050	0.048	0.058
Upstream7	0.074	0.073	0.058	0.011	0.059	0.050	0.049	0.059
Upstream6	0.074	0.072	0.058	0.011	0.059	0.050	0.048	0.058
Upstream5	0.073	0.073	0.058	0.011	0.059	0.050	0.049	0.058
Upstream4	0.074	0.073	0.059	0.011	0.059	0.050	0.049	0.059
Upstream3	0.073	0.073	0.059	0.011	0.059	0.050	0.049	0.059
Upstream2	0.072	0.073	0.061	0.013	0.059	0.050	0.051	0.061
Upstream1	0.067	0.071	0.078	0.032	0.059	0.048	0.066	0.078

A.4 Dinucleotide composition: Quantitative statistics

Anova Results: A comparison of dinucleotide content across the upstream segments

An ANOVA (single factor) analysis was carried out for the occurrence (or proportion) of a given dinucleotide within the ten different upstream datasets; *upstream1*-to-*upstream10*.

Null hypothesis, Ho: the samples of dinucleotide proportion are drawn from the same underlying probability distribution for *upstream1*-to-*upstream10*.

Alternative Hypothesis, H_1: the underlying probability distributions are not the same for all samples. I.e. at least one of the populations (of *upstream1*-to-*upstream10)* has a mean not equal to the others.

The ANOVA (single factor) analysis revealed that the null hypothesis was rejected at the 5% level of significance for fifteen out of the sixteen possible dinucleotides (see the summary table below). Therefore within the ten datasets, *upstream1*-to-*upstream10*, at least one of the datasets was significantly different to the others with respect to these fifteen dinucleotides.

The one exceptional dinucleotide was ApG for which the null hypothesis was accepted. Therefore the underlying distribution was the same for all ten upstream datasets with respect to ApG content only. This means essentially that the ApG compositon is uniform across the ten 1Kb upstream sequence segments.

ANOVA summary table: single factor analysis at 5% level of significance			
For ten upstream datasets: upstream1-to-upstream10			
Dinucleotide	**Probability**	**Ho (reject/accept)**	**datasets -inference**
ApA	0.0000	reject	different
ApT	0.0000	reject	different
ApC	0.0000	reject	different
ApG	0.1469	accept	same
TpA	0.0000	reject	different
TpT	0.0000	reject	different
TpC	0.0000	reject	different
TpG	0.0000	reject	different
CpA	0.0000	reject	different
CpT	0.0000	reject	different
CpC	0.0000	reject	different
CpG	0.0000	reject	different
GpA	0.0014	reject	different
GpT	0.0000	reject	different
GpC	0.0000	reject	different
GpG	0.0000	reject	different

The following data-tables show a more detailed breakdown of the ANOVA results for each of the individual sixteen dinucleotides shown in the summary table above.

ApA: ANOVA Single Factor analysis						
Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	4.14	9.000	0.460	378	0.000	1.8799
Within Groups	227.67	187240	0.001			
Total	231.8	187249				

ApT: ANOVA Single Factor analysis						
Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	4.62	9.000	0.513	911	0.000	1.8799
Within Groups	105.41	187240	0.001			
Total	110.0	187249				

ApC: ANOVA Single Factor analysis						
Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.13	9.000	0.015	123	0.000	1.8799
Within Groups	22.36	187240	0.000			
Total	22.5	187249				

ApG: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.00	9.000	0.000	1	0.147	1.8799
Within Groups	42.33	187240	0.000			
Total	42.3	187249				

TpA: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2.91	9.000	0.323	606	0.000	1.8799
Within Groups	99.99	187240	0.001			
Total	102.9	187249				

TpT: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	4.67	9.000	0.519	435	0.000	1.8799
Within Groups	223.53	187240	0.001			
Total	228.2	187249				

TpC: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.01	9.000	0.001	5	0.000	1.8799
Within Groups	36.05	187240	0.000			
Total	36.1	187249				

TpG: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.85	9.000	0.095	521	0.000	1.8799
Within Groups	34.16	187240	0.000			
Total	35.0	187249				

CpA: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.80	9.000	0.089	518	0.000	1.8799
Within Groups	32.30	187240	0.000			
Total	33.1	187249				

CpT: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.06	9.000	0.007	29	0.000	1.8799
Within Groups	42.94	187240	0.000			
Total	43.0	187249				

CpC: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	6.51	9.000	0.724	910	0.000	1.8799
Within Groups	148.96	187240	0.001			

Total	155.5	187249				

CpG: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	9.80	9.000	1.089	3996	0.000	1.8799
Within Groups	51.02	187240	0.000			
Total	60.8	187249				

GpA: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.01	9.000	0.001	3	0.001	1.8799
Within Groups	36.24	187240	0.000			
Total	36.2	187249				

GpT: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.09	9.000	0.010	86	0.000	1.8799
Within Groups	22.63	187240	0.000			
Total	22.7	187249				

GpC: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	5.85	9.000	0.650	1716	0.000	1.8799
Within Groups	70.90	187240	0.000			
Total	76.8	187249				

GpG: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	6.90	9.000	0.766	974	0.000	1.8799
Within Groups	147.34	187240	0.001			
Total	154.2	187249				

Significance testing: A comparison of dinucleotide content across pairs of adjacent upstream segments

T-tests of the dinucleotide content of adjacent upstream positional segments were used to determine whether the two samples were likely to have come from the same two underlying populations that have the same mean. In other words, *upstream1* was compared with *upstream2*, *upstream2* with *upstream3* etc… It was not assumed that the populations contained an equal variance. These T-tests were two-tailed and carried out at the 5% significance level.

Null hypothesis, Ho: the samples (of dinucleotide proportion) from the two adjacent upstream datasets ($upstream_x$ and $upstream_{x+1}$) are drawn from the same underlying probability distribution.

Alternative Hypothesis, H_1: the underlying probability distributions are not the same for these two positionally adjacent sets of samples

The summary table below shows the results for these T-tests of adjacent segments across the 10Kb upstream. For all sixteen possible dinucleotides the comparison between adjacent pairs of datasets revealed that;

- For 15/16 of the dinucleotides *upstream1* is significantly different to *upstream2*
- For 12/16 *upstream2* is significantly different to *upstream3*
- For 5/16 *upstream3* is significantly different to *upstream4*
- For only 1/16 *upstream4* is significantly different to *upstream5*
- For 2/16 *upstream7* is significantly different to *upstream8*

Therefore for the majority of the dinucleotides differences between the datasets are seen up to *upstream3*. Any differences between adjacent datasets further upstream only occur for a minority of the sixteen possible dinucleotides.

The descriptive statistics showed that there is an increased difference between the upstream segments towards *upstream1*. It therefore appears from these results that the three datasets derived from adjacent upstream regions (*upstream1, upstream2, upstream3*) are different to each other with respect to most dinucleotides and that this is increasingly the case in the direction of the TSS. For the most-part the remaining 7Kb sequence possesses similar dinucleotide content.

SUMMARY TABLE: Comparison of Adjacent upstream segments: two-tailed T-tests at 5% level.
P values highlighted are for pairs of upstream datasets found to be significantly different

Adjacent upstream datasets / Dinucleotides	upstream9/10	upstream8/9	upstream7/8	upstream6/7	upstream5/6	upstream4/5	upstream3/4	upstream2/3	upstream1/2
ApA	0.6503	0.1889	0.0176	0.9294	0.9265	0.0409	0.4690	0.0000	0.0000
ApT	0.7083	0.2995	0.2474	0.8341	0.6237	0.5772	0.0002	0.0000	0.0000
ApC	0.0976	0.6111	0.9255	0.6585	0.1285	0.9203	0.9768	0.5980	0.0000
ApG	0.9259	0.3326	0.8987	0.7919	0.8385	0.7894	0.8619	0.0822	0.0007
TpA	0.5010	0.2411	0.3513	0.8788	0.6384	0.8979	0.0158	0.0000	0.0000
TpT	0.3158	0.8087	0.9240	0.8508	0.4598	0.8163	0.7897	0.0000	0.0000
TpC	0.8512	0.3910	0.0524	0.5937	0.9056	0.2671	0.5334	0.9836	0.0002
TpG	0.7675	0.9189	0.1544	0.4756	0.4950	0.9061	0.3351	0.0000	0.0000
CpA	0.0532	0.4493	0.5556	0.9806	0.3686	0.6098	0.0108	0.0000	0.0000

CpT	0.7636	0.7894	0.0836	0.8589	0.5233	0.4140	0.0614	0.0203	0.0000
CpC	0.8260	0.8194	0.0369	0.6516	0.5355	0.1557	0.1278	0.0000	0.0000
CpG	0.9997	0.4747	0.1861	0.9275	0.8572	0.7439	0.0001	0.0000	0.0000
GpA	0.9191	0.8898	0.3140	0.9896	0.8503	0.5129	0.8155	0.0181	0.1674
GpT	0.7225	0.6018	0.2345	0.7094	0.9198	0.1412	0.9199	0.1314	0.0000
GpC	0.9575	0.3348	0.1747	0.6382	0.3852	0.8020	0.0202	0.0000	0.0000
GpG	0.9938	0.6212	0.5195	0.7075	0.9505	0.2384	0.1195	0.0000	0.0000

The following data-tables show a more detailed breakdown of the T-test results for each of the individual sixteen dinucleotides (ApA, ApT, ApC, ApG, TpA, TpT, TpC, TpG, CpA, CpT, CpC and CpG) that were shown above in the summary table.

ApA

Results for two-tailed T-tests of upstream adjacent datasets:
5% level of significance

	Probability	Ho (reject/accept)	datasets -inference
UPSTREAM9 / upstream10	0.6503	accept	same
UPSTREAM8 / upstream9	0.1889	accept	same
UPSTREAM7 / upstream8	0.0176	reject	different
UPSTREAM6 / upstream7	0.9294	accept	same
UPSTREAM5 / upstream6	0.9265	accept	same
UPSTREAM4 / upstream5	0.0409	reject	different
UPSTREAM3 / upstream4	0.4690	accept	same
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

ApT

Results for two-tailed T-tests of upstream adjacent datasets:
5% level of significance

	Probability	Ho (reject/accept)	datasets -inference
UPSTREAM9 / upstream10	0.7083	accept	same
UPSTREAM8 / upstream9	0.2995	accept	same
UPSTREAM7 / upstream8	0.2474	accept	same
UPSTREAM6 / upstream7	0.8341	accept	same
UPSTREAM5 / upstream6	0.6237	accept	same
UPSTREAM4 / upstream5	0.5772	accept	same
UPSTREAM3 / upstream4	0.0002	reject	different
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

ApC

Results for two-tailed T-tests of upstream adjacent datasets:
5% level of significance

	Probability	Ho (reject/accept)	datasets -inference
UPSTREAM9 / upstream10	0.0976	accept	same
UPSTREAM8 / upstream9	0.6111	accept	same
UPSTREAM7 / upstream8	0.9255	accept	same
UPSTREAM6 / upstream7	0.6585	accept	same
UPSTREAM5 / upstream6	0.1285	accept	same
UPSTREAM4 / upstream5	0.9203	accept	same
UPSTREAM3 / upstream4	0.9768	accept	same
UPSTREAM2 / upstream3	0.5980	accept	same
UPSTREAM1 / upstream2	0.0000	reject	different

ApG

Results for two-tailed T-tests of upstream adjacent datasets:
5% level of significance

	Probability	Ho (reject/accept)	datasets -inference
UPSTREAM9 / upstream10	0.9259	accept	same
UPSTREAM8 / upstream9	0.3326	accept	same
UPSTREAM7 / upstream8	0.8987	accept	same
UPSTREAM6 / upstream7	0.7919	accept	same
UPSTREAM5 / upstream6	0.8385	accept	same
UPSTREAM4 / upstream5	0.7894	accept	same
UPSTREAM3 / upstream4	0.8619	accept	same
UPSTREAM2 / upstream3	0.0822	accept	same
UPSTREAM1 / upstream2	0.0007	reject	different

TpA

Results for two-tailed T-tests of upstream adjacent datasets:
5% level of significance

	Probability	Ho (reject/accept)	datasets -inference
UPSTREAM9 / upstream10	0.5010	accept	same
UPSTREAM8 / upstream9	0.2411	accept	same
UPSTREAM7 / upstream8	0.3513	accept	same
UPSTREAM6 / upstream7	0.8788	accept	same
UPSTREAM5 / upstream6	0.6384	accept	same
UPSTREAM4 / upstream5	0.8979	accept	same
UPSTREAM3 / upstream4	0.0158	reject	different
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

TpT

Results for two-tailed T-tests of upstream adjacent datasets:
5% level of significance

	Probability	Ho (reject/accept)	datasets -inference
UPSTREAM9 / upstream10	0.3158	accept	same
UPSTREAM8 / upstream9	0.8087	accept	same
UPSTREAM7 / upstream8	0.9240	accept	same
UPSTREAM6 / upstream7	0.8508	accept	same
UPSTREAM5 / upstream6	0.4598	accept	same
UPSTREAM4 / upstream5	0.8163	accept	same
UPSTREAM3 / upstream4	0.7897	accept	same
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

TpC

Results for two-tailed T-tests of upstream adjacent datasets:
5% level of significance

	Probability	Ho (reject/accept)	datasets -inference
UPSTREAM9 / upstream10	0.8512	accept	same
UPSTREAM8 / upstream9	0.3910	accept	same
UPSTREAM7 / upstream8	0.0524	accept	same
UPSTREAM6 / upstream7	0.5937	accept	same
UPSTREAM5 / upstream6	0.9056	accept	same
UPSTREAM4 / upstream5	0.2671	accept	same
UPSTREAM3 / upstream4	0.5334	accept	same
UPSTREAM2 / upstream3	0.9836	accept	same
UPSTREAM1 / upstream2	0.0002	reject	different

TpG

Results for two-tailed T-tests of upstream adjacent datasets:
5% level of significance

	Probability	Ho (reject/accept)	datasets -inference
UPSTREAM9 / upstream10	0.7675	accept	same
UPSTREAM8 / upstream9	0.9189	accept	same
UPSTREAM7 / upstream8	0.1544	accept	same
UPSTREAM6 / upstream7	0.4756	accept	same

UPSTREAM5 / upstream6	0.4950	accept	same
UPSTREAM4 / upstream5	0.9061	accept	same
UPSTREAM3 / upstream4	0.3351	accept	same
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

CpA

Results for two-tailed T-tests of upstream adjacent datasets:
5% level of significance

	Probability	Ho (reject/accept)	datasets -inference
UPSTREAM9 / upstream10	0.0532	accept	same
UPSTREAM8 / upstream9	0.4493	accept	same
UPSTREAM7 / upstream8	0.5556	accept	same
UPSTREAM6 / upstream7	0.9806	accept	same
UPSTREAM5 / upstream6	0.3686	accept	same
UPSTREAM4 / upstream5	0.6098	accept	same
UPSTREAM3 / upstream4	0.0108	reject	different
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

CpT

Results for two-tailed T-tests of upstream adjacent datasets:
5% level of significance

	Probability	Ho (reject/accept)	datasets -inference
UPSTREAM9 / upstream10	0.7636	accept	same
UPSTREAM8 / upstream9	0.7894	accept	same
UPSTREAM7 / upstream8	0.0836	accept	same
UPSTREAM6 / upstream7	0.8589	accept	same
UPSTREAM5 / upstream6	0.5233	accept	same
UPSTREAM4 / upstream5	0.4140	accept	same
UPSTREAM3 / upstream4	0.0614	accept	same
UPSTREAM2 / upstream3	0.0203	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

CpC

Results for two-tailed T-tests of upstream adjacent datasets:
5% level of significance

	Probability	Ho (reject/accept)	datasets -inference
UPSTREAM9 / upstream10	0.8260	accept	same
UPSTREAM8 / upstream9	0.8194	accept	same
UPSTREAM7 / upstream8	0.0369	reject	different
UPSTREAM6 / upstream7	0.6516	accept	same
UPSTREAM5 / upstream6	0.5355	accept	same
UPSTREAM4 / upstream5	0.1557	accept	same
UPSTREAM3 / upstream4	0.1278	accept	same
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

CpG

Results for two-tailed T-tests of upstream adjacent datasets:
5% level of significance

	Probability	Ho (reject/accept)	datasets -inference
UPSTREAM9 / upstream10	0.9997	accept	same
UPSTREAM8 / upstream9	0.4747	accept	same
UPSTREAM7 / upstream8	0.1861	accept	same
UPSTREAM6 / upstream7	0.9275	accept	same
UPSTREAM5 / upstream6	0.8572	accept	same
UPSTREAM4 / upstream5	0.7439	accept	same
UPSTREAM3 / upstream4	0.0001	reject	different
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

GpA

Results for two-tailed T-tests of upstream adjacent datasets: 5% level of significance			
	Probability	**Ho (reject/accept)**	**datasets -inference**
UPSTREAM9 / upstream10	0.9191	accept	same
UPSTREAM8 / upstream9	0.8898	accept	same
UPSTREAM7 / upstream8	0.3140	accept	same
UPSTREAM6 / upstream7	0.9896	accept	same
UPSTREAM5 / upstream6	0.8503	accept	same
UPSTREAM4 / upstream5	0.5129	accept	same
UPSTREAM3 / upstream4	0.8155	accept	same
UPSTREAM2 / upstream3	0.0181	reject	different
UPSTREAM1 / upstream2	0.1674	accept	same

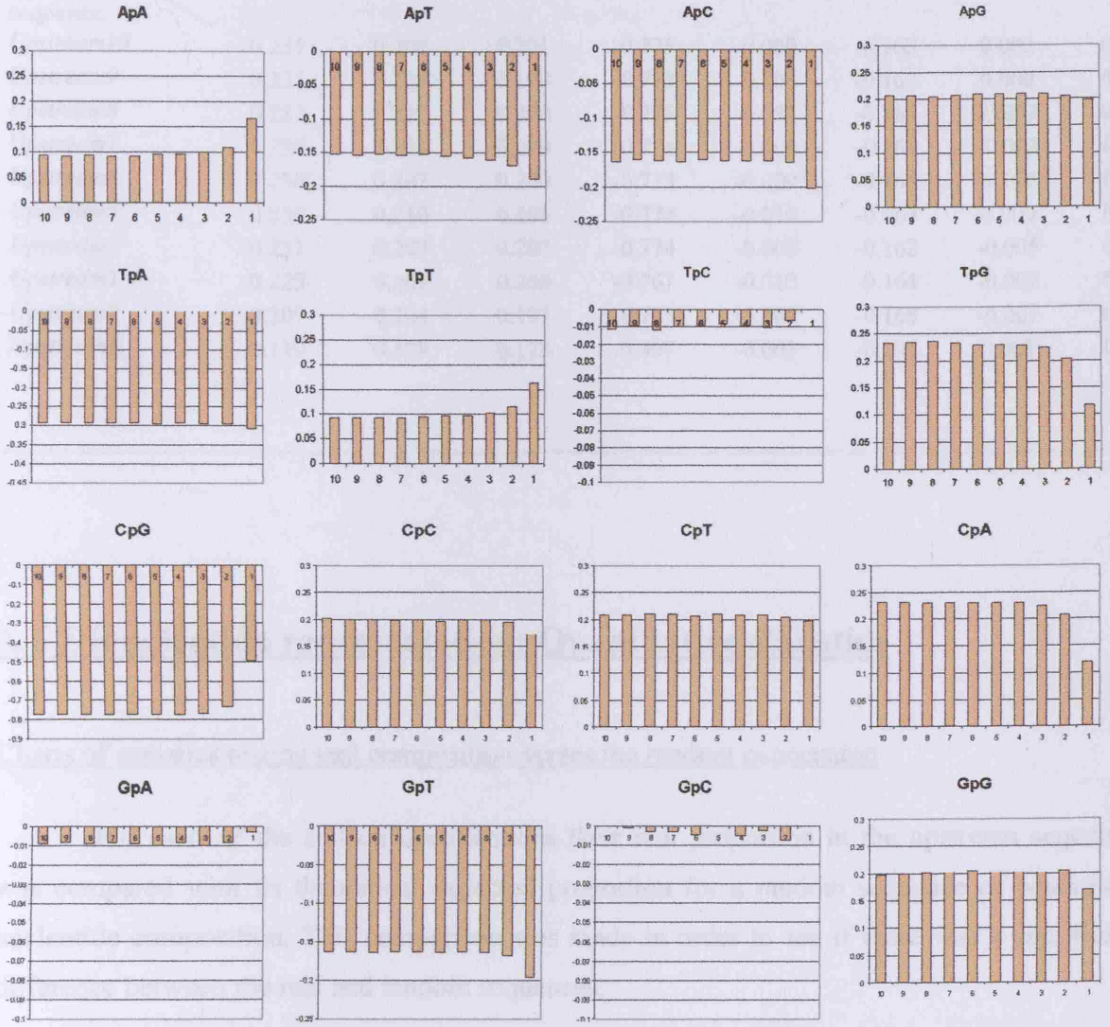
GpT			
Results for two-tailed T-tests of upstream adjacent datasets: 5% level of significance			
	Probability	**Ho (reject/accept)**	**datasets -inference**
UPSTREAM9 / upstream10	0.7225	accept	same
UPSTREAM8 / upstream9	0.6018	accept	same
UPSTREAM7 / upstream8	0.2345	accept	same
UPSTREAM6 / upstream7	0.7094	accept	same
UPSTREAM5 / upstream6	0.9198	accept	same
UPSTREAM4 / upstream5	0.1412	accept	same
UPSTREAM3 / upstream4	0.9199	accept	same
UPSTREAM2 / upstream3	0.1314	accept	same
UPSTREAM1 / upstream2	0.0000	reject	different

GpC			
Results for two-tailed T-tests of upstream adjacent datasets: 5% level of significance			
	Probability	**Ho (reject/accept)**	**datasets -inference**
UPSTREAM9 / upstream10	0.9575	accept	same
UPSTREAM8 / upstream9	0.3348	accept	same
UPSTREAM7 / upstream8	0.1747	accept	same
UPSTREAM6 / upstream7	0.6382	accept	same
UPSTREAM5 / upstream6	0.3852	accept	same
UPSTREAM4 / upstream5	0.8020	accept	same
UPSTREAM3 / upstream4	0.0202	reject	different
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

GpG			
Results for two-tailed T-tests of upstream adjacent datasets: 5% level of significance			
	Probability	**Ho (reject/accept)**	**datasets -inference**
UPSTREAM9 / upstream10	0.9938	accept	same
UPSTREAM8 / upstream9	0.6212	accept	same
UPSTREAM7 / upstream8	0.5195	accept	same
UPSTREAM6 / upstream7	0.7075	accept	same
UPSTREAM5 / upstream6	0.9505	accept	same
UPSTREAM4 / upstream5	0.2384	accept	same
UPSTREAM3 / upstream4	0.1195	accept	Same
UPSTREAM2 / upstream3	0.0000	reject	Different
UPSTREAM1 / upstream2	0.0000	reject	Different

A.5 Dinucleotide representation: Descriptive statistics

The following charts and their associated data-tables show the changes in dinucleotide representation (odds ratio: ρxy = fxy/fxfy) across the 10Kb upstream sequence (as separate datasets: *upstream1*-to-*upstream10*). The results are shown for each of the sixteen individual dinucleotides.



Dinucleotide distance from randomness (odds ratio-1, median) for different upstream positional segments

Upstream sequence \ Dinucleotides	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
Upstream10	0.094	-0.153	-0.163	0.207	-0.292	0.093	-0.011	0.232
Upstream9	0.093	-0.154	-0.161	0.207	-0.292	0.093	-0.009	0.234
Upstream8	0.095	-0.152	-0.162	0.205	-0.291	0.093	-0.012	0.236
Upstream7	0.093	-0.154	-0.163	0.208	-0.292	0.093	-0.010	0.233
Upstream6	0.093	-0.153	-0.161	0.208	-0.292	0.094	-0.009	0.230
Upstream5	0.096	-0.155	-0.162	0.208	-0.293	0.096	-0.008	0.230
Upstream4	0.096	-0.157	-0.162	0.210	-0.292	0.096	-0.009	0.228

Upstream3	0.100	-0.161	-0.161	0.208	-0.294	0.102	-0.011	0.225
Upstream2	0.110	-0.170	-0.164	0.206	-0.295	0.115	-0.006	0.202
Upstream1	0.167	-0.189	-0.202	0.198	-0.308	0.162	0.007	0.117

Dinucleotide distance from randomness (odds ratio-1, median) for different upstream positional segments

Upstream sequence \ Dinucleotides	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
Upstream10	0.231	0.208	0.201	-0.775	-0.009	-0.163	0.000	0.198
Upstream9	0.231	0.208	0.199	-0.773	-0.008	-0.162	0.000	0.199
Upstream8	0.231	0.210	0.200	-0.775	-0.009	-0.164	-0.003	0.202
Upstream7	0.230	0.211	0.199	-0.774	-0.009	-0.163	-0.003	0.203
Upstream6	0.230	0.207	0.200	-0.773	-0.009	-0.163	-0.005	0.204
Upstream5	0.230	0.210	0.196	-0.774	-0.010	-0.165	0.000	0.203
Upstream4	0.231	0.209	0.200	-0.774	-0.008	-0.162	-0.005	0.204
Upstream3	0.225	0.209	0.200	-0.767	-0.010	-0.164	-0.002	0.203
Upstream2	0.209	0.204	0.195	-0.735	-0.008	-0.168	-0.007	0.205
Upstream1	0.119	0.198	0.173	-0.497	-0.003	-0.195	0.003	0.169

A.6 Dinucleotide representation: Quantitative statistics

Charts of statistics testing real composition verses the random expectation

For each of the sixteen dinucleotides their real proportion in the upstream sequence was compared with its theoretical expected proportion for a random sequence of equivalent nucleotide composition. This comparison was made in order to see if there was a significant difference between the real and random sequences.

T-tests were carried for each dinucleotide within each individual upstream dataset comparing the real proportion of that dinucleotide in each sequence sample with the equivalent expected value. The T-tests were two-tailed with no assumption made regarding equal variance and carried out at the 5% level of significance.

Null hypothesis, Ho: The samples of dinucleotide proportion within a given upstream region (such as *upstream1*) are from the same underlying distribution as the randomly expected proportion.

Alternative Hypothesis, H_1: The real and random distribution of dinucleotide proportions are different.

The following summary table shows that the majority of the dinucleotides (13/16) are present at a level that is different to the random expectation in all the upstream segment datasets (*upstream1*-to-*upstream10*). **Three of the dinucleotides are present in the upstream sequence at the random level. These are TpC, GpA and GpC.**

Real verses Random Summary Table: Two-tailed T-tests **at 5% level of significance**
For ten upstream datasets: upstream1-to-upstream10:
P values highlighted are for pairs of upstream datasets found to be significantly different

Adjacent upstream datasets / Dinucleotides	upstream10	upstream9	upstream8	upstream7	upstream6	upstream5	upstream4	upstream3	upstream2	upstream1
ApA	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ApT	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ApC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ApG	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
TpA	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
TpT	0.0000	0.0000	0.0000	0.0134	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
TpC	0.1577	0.3988	0.3593	0.3830	0.5288	0.7825	0.2831	0.1674	0.8778	0.1154
TpG	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CpA	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CpT	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CpC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CpG	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GpA	0.5736	0.2876	0.3300	0.5585	0.7330	0.8402	0.5254	0.4503	0.6910	0.7475
GpT	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GpC	0.5805	0.3614	0.1920	0.3139	0.1123	0.9271	0.1713	0.3263	0.0292	0.4046
GpG	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

The following data-tables show a more detailed breakdown of the T-test results which show whether the dinucleotides content for each upstream segment (*upstream1*-to-*upstream10*) is significantly different to the random expectation at the 5% level of significance.

Upstream1: Comparing Real and (theoretical) Random Distributions for Each Dinucleotide
T-TEST: Two-tailed at 5% level of significance

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
T-TEST	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1154	0.0000
Ho	reject	reject	reject	reject	reject	reject	accept	reject
Distribution	non-random	non-random	non-random	non-random	non-random	non-random	random	non-random
	CpA	**CpT**	**CpC**	**CpG**	**GpA**	**GpT**	**GpC**	**GpG**
T-TEST	0.0000	0.0000	0.0000	0.0000	0.7475	0.0000	0.4046	0.0000
Ho	reject	reject	reject	reject	accept	reject	accept	reject
Distribution	non-	non-	non-	non-	random	non-	random	non-

	random	random	random	random		random		random

Upstream2: Comparing Real and (theoretical) Random Distributions for Each Dinucleotide
T-TEST: Two-tailed at 5% level of significance

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
T-TEST	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.8778	0.0000
Ho	reject	reject	reject	reject	reject	reject	accept	reject
Distribution	non-random	non-random	non-random	non-random	non-random	non-random	random	non-random

	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
T-TEST	0.0000	0.0000	0.0000	0.0000	0.6910	0.0000	0.0292	0.0000
Ho	reject	reject	reject	reject	accept	reject	reject	Reject
Distribution	non-random	non-random	non-random	non-random	random	non-random	non-random	non-random

Upstream3: Comparing Real and (theoretical) Random Distributions for Each Dinucleotide
T-TEST: Two-tailed at 5% level of significance

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
T-TEST	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1674	0.0000
Ho	reject	reject	reject	reject	reject	reject	accept	reject
Distribution	non-random	non-random	non-random	non-random	non-random	non-random	random	non-random

	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
T-TEST	0.0000	0.0000	0.0000	0.0000	0.4503	0.0000	0.3263	0.0000
Ho	rejcct	rejcct	rejcct	rejcct	accept	rejcct	accept	rejcct
Distribution	non-random	non-random	non-random	non-random	random	non-random	random	non-random

Upstream4: Comparing Real and (theoretical) Random Distributions for Each Dinucleotide
T-TEST: Two-tailed at 5% level of significance

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
T-TEST	0.0000	0.0000	0.0000	0.0000	0.0000	0.0280	0.2831	0.0000
Ho	reject	reject	reject	reject	reject	reject	accept	reject
Distribution	non-random	non-random	non-random	non-random	non-random	non-random	random	non-random

	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
T-TEST	0.0000	0.0000	0.0000	0.0000	0.5254	0.0000	0.1713	0.0000
Ho	reject	reject	reject	reject	accept	reject	accept	reject
Distribution	non-random	non-random	non-random	non-random	random	non-random	random	non-random

Upstream5: Comparing Real and (theoretical) Random Distributions for Each Dinucleotide
T-TEST: Two-tailed at 5% level of significance

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
T-TEST	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7825	0.0000
Ho	reject	reject	reject	reject	reject	reject	accept	reject
Distribution	non-random	non-random	non-random	non-random	non-random	non-random	random	non-random

	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
T-TEST	0.0000	0.0000	0.0000	0.0000	0.8402	0.0000	0.9271	0.0000
Ho	reject	reject	reject	reject	accept	reject	accept	reject
Distribution	non-random	non-random	non-random	non-random	random	non-random	random	non-random

Upstream6: Comparing Real and (theoretical) Random Distributions for Each Dinucleotide
T-TEST: Two-tailed at 5% level of significance

	ApT	ApC	ApG	TpA	TpT	TpC	TpG

	ApA							
T-TEST	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.5288	0.0000
Ho	reject	reject	reject	reject	reject	reject	accept	reject
Distribution	non-random	non-random	non-random	non-random	non-random	non-random	random	non-random
	CpA	**CpT**	**CpC**	**CpG**	**GpA**	**GpT**	**GpC**	**GpG**
T-TEST	0.0000	0.0000	0.0000	0.0000	0.7330	0.0000	0.1923	0.0000
Ho	reject	reject	reject	reject	accept	reject	accept	reject
Distribution	non-random	non-random	non-random	non-random	random	non-random	random	non-random

Upstream7: Comparing Real and (theoretical) Random Distributions for Each Dinucleotide
T-TEST: Two-tailed at 5% level of significance

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
T-TEST	0.0000	0.0000	0.0000	0.0000	0.0000	0.0134	0.3830	0.0000
Ho	reject	reject	reject	reject	reject	reject	accept	reject
Distribution	non-random	non-random	non-random	non-random	non-random	non-random	random	non-random
	CpA	**CpT**	**CpC**	**CpG**	**GpA**	**GpT**	**GpC**	**GpG**
T-TEST	0.0000	0.0000	0.0000	0.0000	0.5585	0.0000	0.3139	0.0000
Ho	reject	reject	reject	reject	accept	reject	accept	reject
Distribution	non-random	non-random	non-random	non-random	random	non-random	random	non-random

Upstream8: Comparing Real and (theoretical) Random Distributions for Each Dinucleotide
T-TEST: Two-tailed at 5% level of significance

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
T-TEST	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.3593	0.0000
Ho	reject	reject	reject	reject	reject	reject	accept	reject
Distribution	non-random	non-random	non-random	non-random	non-random	non-random	random	non-random
	CpA	**CpT**	**CpC**	**CpG**	**GpA**	**GpT**	**GpC**	**GpG**
T-TEST	0.0000	0.0000	0.0000	0.0000	0.3300	0.0000	0.1920	0.0000
Ho	reject	reject	reject	reject	accept	reject	accept	reject
Distribution	non-random	non-random	non-random	non-random	random	non-random	random	non-random

Upstream9: Comparing Real and (theoretical) Random Distributions for Each Dinucleotide
T-TEST: Two-tailed at 5% level of significance

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
T-TEST	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.3988	0.0000
Ho	reject	reject	reject	reject	reject	reject	accept	reject
Distribution	non-random	non-random	non-random	non-random	non-random	non-random	random	non-random
	CpA	**CpT**	**CpC**	**CpG**	**GpA**	**GpT**	**GpC**	**GpG**
T-TEST	0.0000	0.0000	0.0000	0.0000	0.2876	0.0000	0.3614	0.0000
Ho	reject	reject	reject	reject	accept	reject	accept	reject
Distribution	non-random	non-random	non-random	non-random	random	non-random	random	non-random

Upstream10: Comparing Real and (theoretical) Random Distributions for Each Dinucleotide
T-TEST: Two-tailed at 5% level of significance

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
T-TEST	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1577	0.0000
Ho	reject	reject	reject	reject	reject	reject	accept	reject
Distribution	non-random	non-random	non-random	non-random	non-random	non-random	random	non-random

	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
T-TEST	0.0000	0.0000	0.0000	0.0000	0.5736	0.0000	0.5805	0.0000
Ho	reject non-	reject non-	reject non-	reject non-	accept	reject non-	accept	reject non-
Distribution	random	random	random	random	random	random	random	random

A.7 Repeats: Proportion of masking in upstream sequence

There are 10 datasets; *upstream1*-to-*upstream10*, spanning 10Kb upstream of the start site of transcription (TSS), *upstream1* being closest to the TSS. Each dataset (*upstream1*-to-*upstream10*) contains 18,725, 1Kb DNA sequence fragments from sequence upstream of 18,725 different mRNA's.

The overall effect of repeats on sequence composition across the 10Kb upstream

These tables show the proportion of masking or presence of repeats in the ten 1Kb upstream sequence datasets; *upstream1*-to-*upstream10*. All of the 18,725, 1Kb DNA sequence fragments from each of these ten datasets was masked for repeats. In the following table is shown the number of upstream fragments (out of 18,725 for each dataset) against the number of repeat masked nucleotides.

Sequence Length –number of nucleotides remaining after Repeat masking	**Number upstream of sequence fragments (out of a dataset of 18,725)**				
	upstream_10	**upstream_9**	**upstream_8**	**upstream_7**	**upstream_6**
x = 0 *	1486	1516	1450	1396	1342
0 > x <= 50	1421	1345	1387	1391	1338
50 > x <= 100	805	784	823	743	720
100 > x <= 150	689	685	678	709	730
150 > x <= 200	649	643	723	693	703
200 > x <= 250	676	706	677	676	698
250 > x <= 300	707	708	646	764	693
300 > x <= 350	734	728	702	724	754
350 > x <= 400	801	736	761	757	778
400 > x <= 450	750	775	759	776	800
450 > x <= 500	781	762	769	795	795
500 > x <= 550	782	783	825	775	754
550 > x <= 600	772	774	709	809	782
600 > x <= 650	743	755	772	758	781
650 > x <= 700	967	964	964	918	1009
700 > x <= 750	771	847	842	803	764
750 > x <= 800	668	693	674	703	678
800 > x <= 850	715	658	659	710	713
850 > x <= 900	646	686	670	687	683

900 > x <= 950	677	698	679	639	666
950 > x <= 1000	2485	2479	2556	2499	2544
****Total number of sequences**	**18725**	**18725**	**18725**	**18725**	**18725**
Mean sequence length	**489.5037**	**492.9266**	**493.7003**	**493.6853**	**498.0185**

Sequence Length –number of nucleotides remaining after Repeat masking	**Number upstream of sequences (out of a dataset of 18,725)**				
	upstream_5	**upstream_4**	**upstream_3**	**upstream_2**	**upstream_1**
***x = 0**	1193	1040	846	547	66
0 > x <= 50	1316	1287	1134	759	135
50 > x <= 100	797	690	669	510	119
100 > x <= 150	727	655	649	562	131
150 > x <= 200	658	668	677	565	169
200 > x <= 250	676	685	683	628	240
250 > x <= 300	727	743	735	650	262
300 > x <= 350	733	726	739	759	342
350 > x <= 400	812	796	814	819	466
400 > x <= 450	798	816	812	790	552
450 > x <= 500	765	803	776	823	545
500 > x <= 550	806	779	802	834	725
550 > x <= 600	816	814	816	862	863
600 > x <= 650	756	823	862	844	896
650 > x <= 700	1014	1035	1011	1067	1281
700 > x <= 750	834	891	870	944	1144
750 > x <= 800	646	689	729	811	1130
800 > x <= 850	663	723	788	850	1347
850 > x <= 900	659	757	775	855	1527
900 > x <= 950	690	710	723	902	1609
950 > x <= 1000	2639	2595	2815	3344	5176
****Total number of sequences**	**18725**	**18725**	**18725**	**18725**	**18725**
Mean sequence length	**503.8203**	**516.4733**	**533.9108**	**583.6993**	**750.4481**

*The sequence is entirely masked (i.e. all repeat)
**The total number of sequences is 18725

A.8 Repeats: Mononucleotides –descriptive statistics

Mononucleotide composition for masked and unmasked upstream sequences								
Upstream positional segment	A		T		C		G	
	repeat masked	unmasked	repeat masked	unmasked	repeat masked	unmasked	repeat masked	unmasked
upstream10	0.283	0.275	0.284	0.276	0.210	0.219	0.210	0.219
upstream9	0.283	0.276	0.284	0.275	0.210	0.219	0.211	0.219
upstream8	0.284	0.276	0.284	0.275	0.210	0.219	0.211	0.219
upstream7	0.282	0.275	0.283	0.275	0.212	0.220	0.211	0.219
upstream6	0.282	0.275	0.284	0.276	0.211	0.220	0.209	0.219
upstream5	0.284	0.276	0.285	0.276	0.210	0.220	0.209	0.219
upstream4	0.284	0.275	0.285	0.276	0.210	0.220	0.211	0.220
upstream3	0.283	0.275	0.284	0.276	0.211	0.221	0.211	0.221
upstream2	0.278	0.272	0.278	0.272	0.217	0.225	0.217	0.225
upstream1	0.234	0.240	0.232	0.237	0.263	0.257	0.262	0.256

Table showing the nucleotide proportions of repeat masked and unmasked upstream sequences, *upstream10*-to-*upstream1*. These proportions are the median values across each dataset.

For both repeat masked and unmasked sequences the proportion of A and T decreases in the downstream direction towards *upstream1* (ie. towards the TSS), while the proportion of C and G increases.

The difference between the masked and unmasked datasets is that in repeat masked sequences the gap between A/T content and C/G content is widened, with A and T being present at relatively higher proportions throughout the upstream and C and G at lower proportions than the unmasked sequences.

Regarding nucleotide proportion trends or changes across the upstream, for both repeat masked and unmasked sequences the proportion of A and T decreases in the downstream direction towards *upstream1*, while the proportion of C and G increases. Also, the A and T proportion is higher than C and G throughout the upstream for both masked and unmasked

sequences, except for *upstream1*. In *upstream1* the C/G content is higher than A/T for both the repeat masked and unmasked datasets.

The difference between repeat masked and unmasked sequences is in the relative gap between A/T content and C/G content. In the repeat masked sequences this gap is widened, with A and T being present at relatively higher proportions throughout the upstream and C and G at lower proportions than the unmasked sequences. For example, within *upstream10* unmasked sequence; C=0.219 and G=0.219. In the repeat masked sequence; C=0.210 and G=0.210. Within *upstream10* unmasked sequence; A=0.275 and T=0.276. In repeat masked sequence; A=0.283 and T=0.284.Within upstream1, this gap is also widened, with the C/G content being higher than A/T in the repeat masked sequence than in the unmasked sequence.

Although there are differences between the masked and unmasked datasets, the overall trend across the 10Kb upstream is the same, with the upstream sequence showing constant nucleotide proportions, except for the 1-3Kb portion (*upstream1*-to-*upstream3*) closest to the start site of transcription. Therefore the apparent changes in nucleotide composition (increase in C/G content and decrease in A/T) in the upstream sequence toward the start site are not due to the presence of repeats.

The following summary charts provide the results of a descriptive statistics analysis for each of mononucleotide content (composition) in the repeat masked 5' upstream region of human genes. There are 10 datasets; *upstream1*-to-*upstream10*, spanning 10Kb upstream of the start site of transcription (TSS), *upstream1* being closest to the TSS. Each dataset (*upstream1*-to-*upstream10*) contains 18,725, 1Kb DNA sequence fragments from sequence upstream of 18,725 different mRNA's.

Upstream1

	A	*T*	*C*	*G*
Mean	0.238	0.237	0.262	0.263
Standard Error	0.0004	0.0004	0.0005	0.0004
Median	0.234	0.232	0.263	0.262
Standard Deviation	0.0575	0.0602	0.0618	0.0606
Skewness	0.268	0.352	0.027	0.134
Count	18404	18404	18404	18404
Confidence Level(95.0%)	0.001	0.001	0.001	0.001

upstream2

	A	*T*	*C*	*G*
Mean	0.238	0.237	0.262	0.263

Standard Error	0.0004	0.0004	0.0005	0.0004
Median	0.234	0.232	0.263	0.262
Standard Deviation	0.0575	0.0602	0.0618	0.0606
Skewness	0.268	0.352	0.027	0.134
Count	18404	18404	18404	18404
Confidence Level(95.0%)	0.001	0.001	0.001	0.001

upstream3

	A	*T*	*C*	*G*
Mean	0.280	0.281	0.220	0.219
Standard Error	0.0005	0.0005	0.0005	0.0005
Median	0.283	0.284	0.211	0.211
Standard Deviation	0.0590	0.0583	0.0582	0.0584
Skewness	-0.089	-0.038	0.539	0.514
Count	16063	16063	16063	16063
Confidence Level(95.0%)	0.001	0.001	0.001	0.001

upstream4

	A	*T*	*C*	*G*
Mean	0.281	0.282	0.219	0.219
Standard Error	0.0005	0.0005	0.0005	0.0005
Median	0.284	0.285	0.210	0.211
Standard Deviation	0.0592	0.0596	0.0586	0.0592
Skewness	-0.082	-0.014	0.561	0.526
Count	15689	15689	15689	15689
Confidence Level(95.0%)	0.001	0.001	0.001	0.001

upstream5

	A	*T*	*C*	*G*
Mean	0.281	0.282	0.219	0.218
Standard Error	0.0005	0.0005	0.0005	0.0005
Median	0.284	0.285	0.210	0.209
Standard Deviation	0.0597	0.0604	0.0592	0.0593
Skewness	-0.066	-0.020	0.585	0.540
Count	15393	15393	15393	15393
Confidence Level(95.0%)	0.001	0.001	0.001	0.001

upstream6

	A	*T*	*C*	*G*
Mean	0.280	0.282	0.219	0.219
Standard Error	0.0005	0.0005	0.0005	0.0005
Median	0.282	0.284	0.211	0.209
Standard Deviation	0.0605	0.0603	0.0593	0.0598
Skewness	0.002	-0.036	0.573	0.553
Count	15309	15309	15309	15309
Confidence Level(95.0%)	0.001	0.001	0.001	0.001

upstream7

	A	*T*	*C*	*G*
Mean	0.280	0.281	0.219	0.219
Standard Error	0.0005	0.0005	0.0005	0.0005
Median	0.282	0.283	0.212	0.211
Standard Deviation	0.0608	0.0604	0.0599	0.0589
Skewness	0.012	-0.014	0.529	0.527
Count	15181	15181	15181	15181

Confidence Level(95.0%)	0.001	0.001	0.001	0.001

upstream8

	A	*T*	*C*	*G*
Mean	0.281	0.281	0.218	0.219
Standard Error	0.0005	0.0005	0.0005	0.0005
Median	0.284	0.284	0.210	0.211
Standard Deviation	0.0609	0.0601	0.0591	0.0590
Skewness	-0.015	0.004	0.528	0.497
Count	15048	15048	15048	15048
Confidence Level(95.0%)	0.001	0.001	0.001	0.001

upstream9

	A	*T*	*C*	*G*
Mean	0.280	0.281	0.218	0.220
Standard Error	0.0005	0.0005	0.0005	0.0005
Median	0.283	0.284	0.210	0.211
Standard Deviation	0.0612	0.0604	0.0593	0.0597
Skewness	0.005	-0.010	0.519	0.544
Count	15062	15062	15062	15062
Confidence Level(95.0%)	0.001	0.001	0.001	0.001

upstream10

	A	*T*	*C*	*G*
Mean	0.280	0.282	0.219	0.219
Standard Error	0.0005	0.0005	0.0005	0.0005
Median	0.283	0.284	0.210	0.210
Standard Deviation	0.0613	0.0610	0.0592	0.0599
Skewness	0.019	-0.016	0.534	0.551
Count	14995	14995	14995	14995
Confidence Level(95.0%)	0.001	0.001	0.001	0.001

The following are charts and data-tables that show the changes in the median and inter-quartile range across the 10Kb upstream for each of the mononucleotides. For the repeat masked sequences, the inter-quartile range decreases (from *upstream10*-to-*upstream1*) towards the start site region. This is the opposite to the trend seen in unmasked sequences where the inter-quartile range increased towards the TSS. This difference though between the masked and unmasked datasets may be due to the reduction in repeat content of the sequence towards the TSS.

Adenine

Upstream segment	upper quartile	lower quartile	median	Inter-quartile range
upstream_10	0.349	0.237	0.283	0.112
upstream_9	0.348	0.237	0.283	0.111

upstream_8	0.348	0.237	0.284	0.111
upstream_7	0.347	0.238	0.282	0.109
upstream_6	0.345	0.237	0.282	0.107
upstream_5	0.343	0.240	0.284	0.103
upstream_4	0.341	0.239	0.284	0.102
upstream_3	0.337	0.238	0.283	0.099
upstream_2	0.326	0.234	0.278	0.092
upstream_1	0.280	0.196	0.234	0.084

Thymine

Upstream segment	upper quartile	lower quartile	median	Inter-quartile range
upstream_10	0.349	0.238	0.284	0.111
upstream_9	0.349	0.238	0.284	0.110
upstream_8	0.348	0.238	0.284	0.110
upstream_7	0.347	0.239	0.283	0.107
upstream_6	0.346	0.239	0.284	0.107
upstream_5	0.345	0.240	0.285	0.105
upstream_4	0.342	0.240	0.285	0.102
upstream_3	0.337	0.240	0.284	0.097
upstream_2	0.327	0.234	0.278	0.093
upstream_1	0.280	0.193	0.232	0.087

Cytosine

Upstream segment	upper quartile	lower quartile	median	Inter-quartile range
upstream_10	0.293	0.174	0.210	0.118
upstream_9	0.292	0.174	0.210	0.117
upstream_8	0.292	0.174	0.210	0.118
upstream_7	0.291	0.174	0.212	0.117
upstream_6	0.288	0.175	0.211	0.114
upstream_5	0.286	0.175	0.210	0.111
upstream_4	0.283	0.175	0.210	0.108
upstream_3	0.279	0.176	0.211	0.103
upstream_2	0.276	0.181	0.217	0.096
upstream_1	0.308	0.217	0.263	0.091

Guanine

Upstream segment	upper quartile	lower quartile	median	Inter-quartile range
upstream_10	0.294	0.175	0.210	0.119
upstream_9	0.293	0.175	0.211	0.118
upstream_8	0.293	0.175	0.211	0.118
upstream_7	0.289	0.174	0.211	0.115
upstream_6	0.289	0.174	0.209	0.115
upstream_5	0.287	0.175	0.209	0.113
upstream_4	0.283	0.174	0.211	0.109
upstream_3	0.279	0.175	0.211	0.104
upstream_2	0.276	0.181	0.217	0.095
upstream_1	0.307	0.218	0.262	0.089

A.9 Repeats: Mononucleotides –quantitative statistics

Anova Results: A comparison of mononucleotide content across the upstream repeat masked segments

An ANOVA (single factor) analysis was carried out for the occurrence (or proportion) of a given mononucleotide within the ten different upstream datasets; *upstream1*-to-*upstream10*.

Null hypothesis, Ho: the samples *(upstream1*-to-*upstream10)* are drawn from the same underlying probability distribution

Alternative Hypothesis, H_1: the underlying probability distributions are not the same for all samples. I.e. at least one of the populations *(upstream1*-to-*upstream10)* has a mean not equal to the others.

The ANOVA (single factor) analysis revealed that the null hypothesis was rejected at the 5% level of significance for each of the four mononucleotides (see the summary table below).Therefore within the ten datasets, *upstream1*-to-*upstream10*, at least one of the datasets was significantly different to the others at the 5% level of significance. The ANOVA result for all four mononucleotides within repeat masked sequences is identical to that seen for the unmasked sequences.

ANOVA summary table: single factor analysis at 5% level of significance			
For ten upstream repeat masked datasets: upstream1-to-upstream10			
Mononucleotide	**Probability**	**Ho (reject/accept)**	**datasets – (same/different)**
Adenine	0.000000	reject	different
Thymine	0.000000	reject	different
Cytosine	0.000000	reject	different
Guanine	0.000000	reject	different

The following data-tables show a more detailed breakdown of the ANOVA results for each of the individual four mononucleotides.

Adenine: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	28.85	9.000	3.206	898	0.000	1.8799
Within Groups	564.23	158023	0.004			
Total	593.1	158032				

Thymine: ANOVA Single Factor analysis

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	31.26	9.000	3.474	967	0.000	1.8799
Within Groups	567.85	158023	0.004			
Total	599.1	158032				

Cytosine: ANOVA Single Factor analysis

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	29.46	9.000	3.273	929	0.000	1.8799
Within Groups	556.82	158023	0.004			
Total	586.3	158032				

Guanine: ANOVA Single Factor analysis

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	30.67	9.000	3.407	968	0.000	1.8799
Within Groups	556.30	158023	0.004			
Total	587.0	158032				

T-test Details: A comparison of mononucleotide content across pairs of adjacent upstream repeat masked segments

T-tests of the mononucleotide content of adjacent upstream positional segments were used to determine whether two samples are likely to have come from the same two underlying populations that have the same mean. In other words, *upstream1* was compared with *upstream2*, *upstream2* with *upstream3* etc...It was not assumed that the populations contained an equal variance. These T-tests were two-tailed and carried out at the 5% significance level.

Null hypothesis, Ho: the samples (of mononucleotide proportion) from the two adjacent upstream datasets ($upstream_x$ and $upstream_{x+1}$) are drawn from the same underlying probability distribution.

Alternative Hypothesis, H_1: the underlying probability distributions are not the same for these two positionally adjacent sets of samples

The summary table below shows the results for these T-tests of adjacent segments across the 10Kb upstream. For all four mononucleotides the comparison between pairs of datasets revealed that;

- *upstream1* and *upstream2* are significantly different.
- *upstream2* and *upstream3* are significantly different.

The remaining further upstream adjacent pairs of datasets were all found to be significantly different to each other at the 5% level of significance. The remaining datasets spanned a total of 7Kb.

This result for the repeat masked sequences is slightly different to that of the unmasked sequences. In the unmasked sequences there were some additional differences further upstream. For instance, for Adenine in unmasked sequences, a significant difference was found between *upstream3* and *upstream4* and also between *upstream7* and *upstream8*. Therefore, it seems that masking the sequences has resulted in a limiting of significant differences found to *upstream1/upstream2* and *upstream2/upstream3*, showing that the boundary of mononucleotide difference across the segments is up to *upstream3*.

SUMMARY TABLE: Comparison of adjacent *repeat masked* upstream segments: two-tailed T-tests at 5% level.
P values highlighted are for pairs of upstream datasets found to be significantly different

Adjacent upstream datasets / Dinucleotides	Upstream9/ 10	upstream8/ 9	upstream7/ 8	upstream6/ 7	upstream5/ 6	upstream4/ 5	upstream3/ 4	upstream2/ 3	upstream1/ 2
A	0.9238	0.6857	0.5823	0.9083	0.5016	0.9088	0.2597	0.0000	0.0000
T	0.9194	0.7606	0.9696	0.5677	0.7376	0.8552	0.3170	0.0000	0.0000
C	0.8843	0.9868	0.1377	0.6813	0.6624	0.8434	0.1898	0.0000	0.0000
G	0.8817	0.9281	0.3730	0.9596	0.5595	0.8992	0.4041	0.0000	0.0000

The following data-tables show a more detailed breakdown of the T-test results for each of the individual four mononucleotides.

Adenine			
Results for two-tailed T-tests of upstream adjacent datasets: 5% level of significance			
	Probability	**Ho (reject/accept)**	**datasets - same/different)**
UPSTREAM9 / upstream10	0.9238	accept	same
UPSTREAM8 / upstream9	0.6857	accept	same
UPSTREAM7 / upstream8	0.5823	accept	same
UPSTREAM6 / upstream7	0.9083	accept	same
UPSTREAM5 / upstream6	0.5016	accept	same
UPSTREAM4 / upstream5	0.9088	accept	same
UPSTREAM3 / upstream4	0.2597	accept	same
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

Thymine			
Results for two-tailed T-tests of adjacent upstream datasets: 5% level of significance			
	Probability	**Ho (reject/accept)**	**datasets - same/different)**
UPSTREAM9 / upstream10	0.9194	accept	same
UPSTREAM8 / upstream9	0.7606	accept	same
UPSTREAM7 / upstream8	0.9696	accept	same
UPSTREAM6 / upstream7	0.5677	accept	same
UPSTREAM5 / upstream6	0.7376	accept	same
UPSTREAM4 / upstream5	0.8552	accept	same
UPSTREAM3 / upstream4	0.3170	accept	same
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

Cytosine			
Results for two-tailed T-tests of adjacent upstream datasets 5% level of significance			
	Probability	**Ho (reject/accept)**	**datasets - same/different)**
UPSTREAM9 / upstream10	0.8843	accept	same
UPSTREAM8 / upstream9	0.9868	accept	same
UPSTREAM7 / upstream8	0.1377	accept	same
UPSTREAM6 / upstream7	0.6813	accept	same
UPSTREAM5 / upstream6	0.6624	accept	same
UPSTREAM4 / upstream5	0.8434	accept	same
UPSTREAM3 / upstream4	0.1898	accept	same
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

Guanine			
Results for two-tailed T-tests of adjacent upstream datasets 5% level of significance			
	Probability	**Ho (reject/accept)**	**datasets - same/different)**
UPSTREAM9 / upstream10	0.8817	accept	same
UPSTREAM8 / upstream9	0.9281	accept	same
UPSTREAM7 / upstream8	0.3730	accept	same
UPSTREAM6 / upstream7	0.9596	accept	same
UPSTREAM5 / upstream6	0.5595	accept	same
UPSTREAM4 / upstream5	0.8992	accept	same
UPSTREAM3 / upstream4	0.4041	accept	same
UPSTREAM2 / upstream3	0.0000	reject	different

UPSTREAM1 / upstream2	0.0000	reject	different

A.10 Repeats: Dinucleotide composition –descriptive statistics

The following summary charts provide the results of a descriptive statistics analysis for each of dinucleotide content (composition) in the repeat masked 5' upstream region of human genes. There are 10 datasets; *upstream1*-to-*upstream10*, spanning 10Kb upstream of the start site of transcription (TSS), *upstream1* being closest to the TSS. Each dataset (*upstream1*-to-*upstream10*) contains 18,725, 1Kb DNA sequence fragments from sequence upstream of 18,725 different mRNA's.

Upstream1

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
Mean	0.069	0.048	0.047	0.073	0.041	0.069	0.061	0.066
Standard Error	0.0002	0.0002	0.0001	0.0001	0.0002	0.0003	0.0001	0.0001
Median	0.065	0.043	0.047	0.072	0.036	0.063	0.060	0.065
Standard Deviation	0.0324	0.0251	0.0110	0.0149	0.0237	0.0339	0.0135	0.0156
Skewness	0.645	0.754	0.586	0.434	0.851	0.742	0.449	0.644
Count	18404	18404	18404	18404	18404	18404	18404	18404
Confidence Level(95.0%)	0.0005	0.0004	0.0002	0.0002	0.0003	0.0005	0.0002	0.0002

Upstream1

	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
Mean	0.066	0.072	0.083	0.039	0.061	0.048	0.070	0.084
Standard Error	0.0001	0.0001	0.0003	0.0002	0.0001	0.0001	0.0002	0.0003
Median	0.066	0.071	0.081	0.034	0.060	0.047	0.068	0.081
Standard Deviation	0.0137	0.0148	0.0359	0.0306	0.0137	0.0113	0.0285	0.0363
Skewness	0.437	0.400	0.374	0.710	0.590	0.901	0.400	0.451
Count	18404	18404	18404	18404	18404	18404	18404	18404
Confidence Level(95.0%)	0.0002	0.0002	0.0005	0.0004	0.0002	0.0002	0.0004	0.0005

Upstream2

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
Mean	0.087	0.065	0.050	0.073	0.056	0.087	0.061	0.071
Standard Error	0.0003	0.0002	0.0001	0.0001	0.0002	0.0003	0.0001	0.0001
Median	0.086	0.065	0.049	0.072	0.055	0.085	0.060	0.070
Standard Deviation	0.0366	0.0265	0.0124	0.0178	0.0269	0.0374	0.0160	0.0157
Skewness	0.377	0.222	0.795	0.391	0.304	0.441	0.446	0.492
Count	16889	16889	16889	16889	16889	16889	16889	16889
Confidence Level(95.0%)	0.0006	0.0004	0.0002	0.0003	0.0004	0.0006	0.0002	0.0002

Upstream2

	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
Mean	0.071	0.073	0.064	0.016	0.060	0.050	0.049	0.064
Standard Error	0.0001	0.0001	0.0003	0.0001	0.0001	0.0001	0.0002	0.0003
Median	0.071	0.072	0.057	0.010	0.060	0.049	0.044	0.057
Standard Deviation	0.0149	0.0176	0.0336	0.0178	0.0161	0.0123	0.0224	0.0335
Skewness	0.461	0.253	0.921	2.592	0.591	0.951	0.988	0.877
Count	16889	16889	16889	16889	16889	16889	16889	16889
Confidence Level(95.0%)	0.0002	0.0003	0.0005	0.0003	0.0002	0.0002	0.0003	0.0005

Upstream3

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
Mean	0.089	0.068	0.050	0.072	0.058	0.089	0.061	0.072
Standard Error	0.0003	0.0002	0.0001	0.0001	0.0002	0.0003	0.0001	0.0001
Median	0.087	0.068	0.049	0.071	0.058	0.088	0.060	0.071
Standard Deviation	0.0379	0.0271	0.0127	0.0184	0.0276	0.0377	0.0167	0.0157
Skewness	0.363	0.162	1.030	0.290	0.250	0.404	0.525	0.432
Count	16063	16063	16063	16063	16063	16063	16063	16063
Confidence Level(95.0%)	0.0006	0.0004	0.0002	0.0003	0.0004	0.0006	0.0003	0.0002

Upstream3

	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
Mean	0.072	0.073	0.061	0.013	0.060	0.050	0.046	0.061
Standard Error	0.0001	0.0001	0.0003	0.0001	0.0001	0.0001	0.0002	0.0003
Median	0.071	0.072	0.054	0.008	0.059	0.049	0.042	0.054
Standard Deviation	0.0156	0.0185	0.0332	0.0152	0.0167	0.0126	0.0216	0.0329
Skewness	0.630	0.302	0.936	3.068	0.510	0.882	0.960	0.915
Count	16063	16063	16063	16063	16063	16063	16063	16063
Confidence Level(95.0%)	0.0002	0.0003	0.0005	0.0002	0.0003	0.0002	0.0003	0.0005

Upstream4

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
Mean	0.089	0.068	0.050	0.072	0.059	0.089	0.061	0.072
Standard Error	0.0003	0.0002	0.0001	0.0001	0.0002	0.0003	0.0001	0.0001
Median	0.087	0.069	0.049	0.072	0.058	0.088	0.060	0.072
Standard Deviation	0.0381	0.0277	0.0126	0.0186	0.0284	0.0387	0.0165	0.0159
Skewness	0.356	0.137	0.865	0.329	0.254	0.440	0.478	0.493
Count	15689	15689	15689	15689	15689	15689	15689	15689
Confidence Level(95.0%)	0.0006	0.0004	0.0002	0.0003	0.0004	0.0006	0.0003	0.0002

Upstream4

	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
Mean	0.072	0.073	0.061	0.012	0.060	0.050	0.046	0.061
Standard Error	0.0001	0.0001	0.0003	0.0001	0.0001	0.0001	0.0002	0.0003
Median	0.071	0.072	0.053	0.008	0.059	0.049	0.042	0.054
Standard Deviation	0.0156	0.0184	0.0333	0.0149	0.0166	0.0127	0.0216	0.0332
Skewness	0.403	0.305	0.979	3.028	0.551	0.966	0.948	0.936
Count	15689	15689	15689	15689	15689	15689	15689	15689
Confidence Level(95.0%)	0.0002	0.0003	0.0005	0.0002	0.0003	0.0002	0.0003	0.0005

Upstream5

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
Mean	0.089	0.068	0.050	0.072	0.059	0.089	0.061	0.072
Standard Error	0.0003	0.0002	0.0001	0.0002	0.0002	0.0003	0.0001	0.0001
Median	0.088	0.069	0.049	0.072	0.058	0.088	0.060	0.071
Standard Deviation	0.0386	0.0276	0.0127	0.0188	0.0282	0.0392	0.0168	0.0160
Skewness	0.433	0.184	0.850	0.321	0.242	0.464	0.482	0.501
Count	15393	15393	15393	15393	15393	15393	15393	15393
Confidence Level(95.0%)	0.0006	0.0004	0.0002	0.0003	0.0004	0.0006	0.0003	0.0003

Upstream5

	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
Mean	0.072	0.073	0.061	0.012	0.060	0.050	0.047	0.060
Standard Error	0.0001	0.0001	0.0003	0.0001	0.0001	0.0001	0.0002	0.0003
Median	0.071	0.072	0.053	0.008	0.059	0.049	0.042	0.053
Standard Deviation	0.0159	0.0185	0.0337	0.0154	0.0168	0.0128	0.0221	0.0332
Skewness	0.418	0.276	1.042	3.186	0.515	0.947	1.007	0.924
Count	15393	15393	15393	15393	15393	15393	15393	15393
Confidence Level(95.0%)	0.0003	0.0003	0.0005	0.0002	0.0003	0.0002	0.0003	0.0005

Upstream6

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
Mean	0.088	0.068	0.050	0.072	0.059	0.089	0.061	0.072
Standard Error	0.0003	0.0002	0.0001	0.0002	0.0002	0.0003	0.0001	0.0001
Median	0.086	0.069	0.049	0.072	0.059	0.088	0.060	0.072
Standard Deviation	0.0390	0.0279	0.0127	0.0187	0.0286	0.0390	0.0168	0.0161
Skewness	0.481	0.165	0.736	0.325	0.287	0.481	0.484	0.356
Count	15309	15309	15309	15309	15309	15309	15309	15309
Confidence Level(95.0%)	0.0006	0.0004	0.0002	0.0003	0.0005	0.0006	0.0003	0.0003

Upstream6

	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
Mean	0.072	0.073	0.061	0.012	0.060	0.050	0.047	0.061
Standard Error	0.0001	0.0002	0.0003	0.0001	0.0001	0.0001	0.0002	0.0003
Median	0.071	0.072	0.053	0.008	0.059	0.049	0.042	0.053
Standard Deviation	0.0157	0.0186	0.0338	0.0153	0.0169	0.0128	0.0221	0.0339
Skewness	0.448	0.303	1.012	3.148	0.517	0.800	0.982	1.004
Count	15309	15309	15309	15309	15309	15309	15309	15309
Confidence Level(95.0%)	0.0002	0.0003	0.0005	0.0002	0.0003	0.0002	0.0003	0.0005

Upstream7

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
Mean	0.088	0.068	0.050	0.072	0.059	0.089	0.060	0.072
Standard Error	0.0003	0.0002	0.0001	0.0001	0.0002	0.0003	0.0001	0.0001
Median	0.087	0.069	0.049	0.072	0.058	0.087	0.059	0.072
Standard Deviation	0.0393	0.0278	0.0128	0.0184	0.0286	0.0390	0.0170	0.0159
Skewness	0.494	0.174	0.935	0.228	0.304	0.454	0.476	0.312
Count	15181	15181	15181	15181	15181	15181	15181	15181
Confidence Level(95.0%)	0.0006	0.0004	0.0002	0.0003	0.0005	0.0006	0.0003	0.0003

Upstream7

	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
Mean	0.072	0.073	0.061	0.012	0.060	0.050	0.047	0.061
Standard Error	0.0001	0.0002	0.0003	0.0001	0.0001	0.0001	0.0002	0.0003

Median	0.071	0.072	0.054	0.008	0.059	0.049	0.042	0.053
Standard Deviation	0.0159	0.0188	0.0341	0.0154	0.0165	0.0127	0.0221	0.0334
Skewness	0.443	0.280	0.987	3.221	0.381	0.834	0.960	0.938
Count	15181	15181	15181	15181	15181	15181	15181	15181
Confidence Level(95.0%)	0.0003	0.0003	0.0005	0.0002	0.0003	0.0002	0.0004	0.0005

Upstream8

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
Mean	0.089	0.068	0.050	0.072	0.058	0.089	0.060	0.073
Standard Error	0.0003	0.0002	0.0001	0.0002	0.0002	0.0003	0.0001	0.0001
Median	0.088	0.069	0.049	0.072	0.058	0.087	0.059	0.072
Standard Deviation	0.0394	0.0279	0.0128	0.0185	0.0287	0.0393	0.0166	0.0161
Skewness	0.463	0.147	0.905	0.323	0.281	0.503	0.431	0.413
Count	15048	15048	15048	15048	15048	15048	15048	15048
Confidence Level(95.0%)	0.0006	0.0004	0.0002	0.0003	0.0005	0.0006	0.0003	0.0003

Upstream8

	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
Mean	0.072	0.073	0.060	0.012	0.061	0.050	0.047	0.061
Standard Error	0.0001	0.0002	0.0003	0.0001	0.0001	0.0001	0.0002	0.0003
Median	0.072	0.072	0.053	0.008	0.060	0.049	0.042	0.054
Standard Deviation	0.0158	0.0185	0.0334	0.0150	0.0166	0.0130	0.0221	0.0331
Skewness	0.387	0.225	0.919	3.108	0.431	0.871	0.987	0.884
Count	15048	15048	15048	15048	15048	15048	15048	15048
Confidence Level(95.0%)	0.0003	0.0003	0.0005	0.0002	0.0003	0.0002	0.0004	0.0005

Upstream9

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
Mean	0.089	0.068	0.050	0.072	0.058	0.089	0.060	0.073
Standard Error	0.0003	0.0002	0.0001	0.0002	0.0002	0.0003	0.0001	0.0001
Median	0.087	0.068	0.049	0.072	0.058	0.087	0.059	0.072
Standard Deviation	0.0395	0.0280	0.0127	0.0184	0.0289	0.0394	0.0167	0.0162
Skewness	0.446	0.170	0.922	0.307	0.290	0.479	0.466	0.545
Count	15062	15062	15062	15062	15062	15062	15062	15062
Confidence Level(95.0%)	0.0006	0.0004	0.0002	0.0003	0.0005	0.0006	0.0003	0.0003

Upstream9

	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
Mean	0.072	0.073	0.060	0.012	0.060	0.050	0.047	0.061
Standard Error	0.0001	0.0002	0.0003	0.0001	0.0001	0.0001	0.0002	0.0003
Median	0.071	0.072	0.053	0.008	0.060	0.049	0.042	0.053
Standard Deviation	0.0157	0.0184	0.0336	0.0150	0.0167	0.0129	0.0223	0.0336
Skewness	0.472	0.235	0.926	3.080	0.470	1.056	0.944	0.928
Count	15062	15062	15062	15062	15062	15062	15062	15062
Confidence Level(95.0%)	0.0003	0.0003	0.0005	0.0002	0.0003	0.0002	0.0004	0.0005

Upstream10

	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
Mean	0.089	0.068	0.050	0.072	0.058	0.089	0.060	0.072
Standard Error	0.0003	0.0002	0.0001	0.0002	0.0002	0.0003	0.0001	0.0001
Median	0.087	0.069	0.049	0.072	0.058	0.088	0.059	0.072
Standard Deviation	0.0398	0.0280	0.0128	0.0187	0.0286	0.0395	0.0166	0.0161
Skewness	0.518	0.142	0.952	0.409	0.270	0.458	0.510	0.452

Count	14995	14995	14995	14995	14995	14995	14995	14995
Confidence Level(95.0%)	0.0006	0.0004	0.0002	0.0003	0.0005	0.0006	0.0003	0.0003

Upstream10								
	CpA	**CpT**	**CpC**	**CpG**	**GpA**	**GpT**	**GpC**	**GpG**
Mean	0.072	0.073	0.060	0.013	0.060	0.050	0.047	0.061
Standard Error	0.0001	0.0002	0.0003	0.0001	0.0001	0.0001	0.0002	0.0003
Median	0.071	0.072	0.053	0.008	0.059	0.049	0.042	0.053
Standard Deviation	0.0158	0.0186	0.0334	0.0155	0.0167	0.0128	0.0224	0.0338
Skewness	0.440	0.330	0.924	3.213	0.648	0.798	1.012	0.936
Count	14995	14995	14995	14995	14995	14995	14995	14995
Confidence Level(95.0%)	0.0003	0.0003	0.0005	0.0002	0.0003	0.0002	0.0004	0.0005

The following charts and their associated data-tables show the changes in dinucleotide proportion across the 10Kb upstream repeat masked sequence (as separate datasets: *upstream1*-to-*upstream10*). The results are shown for each of the sixteen individual dinucleotides.

Dinucleotide proportions for different upstream positional segments

Upstream sequence \ Dinucleotides	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
upstream_10	0.087	0.069	0.049	0.072	0.058	0.088	0.059	0.072
upstream_9	0.087	0.068	0.049	0.072	0.058	0.087	0.059	0.072
upstream_8	0.088	0.069	0.049	0.072	0.058	0.087	0.059	0.072
upstream_7	0.087	0.069	0.049	0.072	0.058	0.087	0.059	0.072
upstream_6	0.086	0.069	0.049	0.072	0.059	0.088	0.060	0.072
upstream_5	0.088	0.069	0.049	0.072	0.058	0.088	0.060	0.071
upstream_4	0.087	0.069	0.049	0.072	0.058	0.088	0.060	0.072
upstream_3	0.087	0.068	0.049	0.071	0.058	0.088	0.060	0.071
upstream_2	0.086	0.065	0.049	0.072	0.055	0.085	0.060	0.070
upstream_1	0.065	0.043	0.047	0.072	0.036	0.063	0.060	0.065

Dinucleotide proportions for different upstream positional segments

Upstream sequence \ Dinucleotides	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
upstream_10	0.071	0.072	0.053	0.008	0.059	0.049	0.042	0.053

upstream_9	0.071	0.072	0.053	0.008	0.060	0.049	0.042	0.053
upstream_8	0.072	0.072	0.053	0.008	0.060	0.049	0.042	0.054
upstream_7	0.071	0.072	0.054	0.008	0.059	0.049	0.042	0.053
upstream_6	0.071	0.072	0.053	0.008	0.059	0.049	0.042	0.053
upstream_5	0.071	0.072	0.053	0.008	0.059	0.049	0.042	0.053
upstream_4	0.071	0.072	0.053	0.008	0.059	0.049	0.042	0.054
upstream_3	0.071	0.072	0.054	0.008	0.059	0.049	0.042	0.054
upstream_2	0.071	0.072	0.057	0.010	0.060	0.049	0.044	0.057
upstream_1	0.066	0.071	0.081	0.034	0.060	0.047	0.068	0.081

A.11 Repeats: Dinucleotide composition –quantitative statistics

Anova Results: A comparison of dinucleotide content across the upstream repeat masked segments

An ANOVA (single factor) analysis was carried out for the occurrence (or proportion) of a given dinucleotide within the ten different upstream datasets; *upstream1*-to-*upstream10*.

Null hypothesis, Ho: the samples of dinucleotide proportion are drawn from the same underlying probability distribution for *upstream1*-to-*upstream10*.

Alternative Hypothesis, H_1: the underlying probability distributions are not the same for all samples. I.e. at least one of the populations has a mean not equal to the others.

The ANOVA (single factor) analysis revealed that the null hypothesis was rejected at the 5% level of significance for all of the sixteen possible dinucleotides (see the summary table below). Therefore within the ten datasets, *upstream1*-to-*upstream10*, at least one of the datasets was significantly different to the others with respect to all sixteen dinucleotides.

The difference between these ANOVA's for repeat masked sequences and those of unmasked sequences is only for one dinucleotide, ApG. In the unmasked upstream datasets the ApG proportion was found to be the same across the ten datasets. The results for all the other dinucleotides were synonymous between masked and unmasked datasets.

ANOVA summary table: single factor analysis at 5% level of significance

For ten upstream repeat masked datasets: upstream1-to-upstream10			
Dinucleotide	**Probability**	**Ho (reject/accept)**	**datasets - same/different)**
ApA	0.000000	reject	different
ApT	0.000000	reject	different
ApC	3.6E-123	reject	different
ApG	0.007109	reject	different
TpA	0.000000	reject	different
TpT	0.000000	reject	different
TpC	0.000104	reject	different
TpG	0.000000	reject	different
CpA	0.000000	reject	different
CpT	3.53E-11	reject	different
CpC	0.000000	reject	different
CpG	0.000000	reject	different
GpA	0.007776	reject	different
GpT	3.75E-88	reject	different
GpC	0.000000	reject	different
GpG	0.000000	reject	different

The following data-tables show a more detailed breakdown of the ANOVA results for each of the individual sixteen dinucleotides.

ApA: ANOVA Single Factor analysis						
Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	6.15	9.000	0.683	473	0.000	1.8799
Within Groups	227.92	158013	0.001			
Total	234.1	158022				

ApT: ANOVA Single Factor analysis						
Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	6.70	9.000	0.745	999	0.000	1.8799
Within Groups	117.84	158013	0.001			
Total	124.5	158022				

ApC: ANOVA Single Factor analysis						
Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.09	9.000	0.010	67	0.000	1.8799
Within Groups	24.62	158013	0.000			
Total	24.7	158022				

ApG: ANOVA Single Factor analysis						
Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.01	9.000	0.001	3	0.007	1.8799
Within Groups	51.68	158013	0.000			

Total	51.7	158022

TpA: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	4.88	9.000	0.542	703	0.000	1.8799
Within Groups	121.84	158013	0.001			
Total	126.7	158022				

TpT: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	6.47	9.000	0.719	492	0.000	1.8799
Within Groups	231.03	158013	0.001			
Total	237.5	158022				

TpC: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.01	9.000	0.001	4	0.000	1.8799
Within Groups	41.98	158013	0.000			
Total	42.0	158022				

TpG: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.64	9.000	0.071	281	0.000	1.8799
Within Groups	40.01	158013	0.000			
Total	40.6	158022				

CpA: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.48	9.000	0.053	222	0.000	1.8799
Within Groups	37.71	158013	0.000			
Total	38.2	158022				

CpT: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.02	9.000	0.002	8	0.000	1.8799
Within Groups	51.45	158022	0.000			
Total	51.5	158031				

CpC: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	16514894	9.000	1834988	1679	0.000	1.8799
Within Groups	2.05E+08	187250	1093			
Total	2.21E+08	187259				

CpG: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	11.66	9.000	1.295	4021	0.000	1.8799

Within Groups	50.90	158022	0.000			
Total	62.6	158031				

GpA: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.01	9.000	0.001	2	0.008	1.8799
Within Groups	42.06	158022	0.000			
Total	42.1	158031				

GpT: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.07	9.000	0.008	48	0.000	1.8799
Within Groups	24.96	158022	0.000			
Total	25.0	158031				

GpC: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	8.43	9.000	0.937	1786	0.000	1.8799
Within Groups	82.88	158022	0.001			
Total	91.3	158031				

GpG: ANOVA Single Factor analysis

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	8.42	9.000	0.936	821	0.000	1.8799
Within Groups	180.09	158022	0.001			
Total	188.5	158031				

T-test Details: A comparison of dinucleotide content across pairs of adjacent repeat masked upstream segments

T-tests of the dinucleotide content of adjacent upstream positional segments were used to determine whether the two samples were likely to have come from the same two underlying populations that have the same mean. In other words, *upstream1* was compared with *upstream2*, *upstream2* with *upstream3* etc... It was not assumed that the populations contained an equal variance. These T-tests were two-tailed and carried out at the 5% significance level.

Null hypothesis, Ho: the samples (of dinucleotide proportion) from the two adjacent upstream datasets ($upstream_x$ and $upstream_{x+1}$) are drawn from the same underlying probability distribution.

Alternative Hypothesis, H_1: the underlying probability distributions of the positionally adjacent datasets for dinucleotide proportion are not the same for these two sets of samples.

The summary table below shows the results for these T-tests of adjacent segments across the 10Kb upstream. For all sixteen possible dinucleotides the comparison between adjacent pairs of datasets revealed that;

- For 14/16 of the dinucleotides *upstream1* is significantly different to *upstream2*
- For 11/16 *upstream2* is significantly different to *upstream3*
- For only 2/16 *upstream3* is significantly different to *upstream4*

Therefore for the majority of the dinucleotides differences between the datasets are seen up to *upstream3*. For very few dinucleotides (only two out of sixteen) a significant difference is seen between the *upstream3* and *upstream4* datasets. For the remaining further upstream adjacent sequence comparison no significant difference exists between the datasets, with the exception of *upstream7/upstream8* for which two dinucleotides were significantly different in proportions between these datasets.

In general the results for the masked and unmasked datasets were similar. The main difference between then, was that in the masked sequence, significant differences between adjacent upstream segments, were more limited to the first three segments; *upstream1*, *upstream2* and *upstream3*. In other words in the unmasked sequence significant differences were also found further upstream. Therefore as with the mononucleotides we see that masking limits somewhat significant differences to more downstream regions.

SUMMARY TABLE: Comparison of Adjacent upstream segments: two-tailed T-tests at 5% level.
P values highlighted are for pairs of upstream datasets found to be significantly different

Adjacent upstream datasets / Dinucleotides	upstream9/10	upstream8/9	upstream7/8	upstream6/7	upstream5/6	upstream4/5	upstream3/4	upstream2/3	upstream1/2
ApA	0.9599	0.7509	0.5426	0.6380	0.2690	0.8756	0.7838	0.0000	0.0000
ApT	0.8356	0.7554	0.7516	0.6211	0.6792	0.5469	0.0259	0.0000	0.0000
ApC	0.9350	0.6427	0.2148	0.8497	0.5454	0.5455	0.0400	0.3398	0.0000
ApG	0.3811	0.8579	0.3419	0.9244	0.2945	0.5660	0.3204	0.0158	0.1135
TpA	0.9249	0.9395	0.6898	0.6843	0.7342	0.5151	0.0660	0.0000	0.0000
TpT	0.7247	0.5954	0.9802	0.5337	0.5043	0.9459	0.4899	0.0000	0.0000
TpC	0.9036	0.9916	0.1936	0.4273	0.9239	0.5641	0.5864	0.9027	0.1472

TpG	0.7158	0.7122	0.0249	0.3846	0.8287	0.3406	0.5468	0.0000	0.0000
CpA	0.8333	0.4896	0.4951	0.9889	0.7266	0.7296	0.7634	0.0001	0.0000
CpT	0.7470	0.8033	0.4840	0.8019	0.7387	0.7604	0.1372	0.1344	0.0000
CpC	0.9279	0.9955	0.1172	0.4102	0.6916	0.8679	0.4259	0.0000	0.0000
CpG	0.6156	0.6119	0.5353	0.9880	0.3980	0.7051	0.0541	0.0000	0.0000
GpA	0.4921	0.6739	0.0380	0.9813	0.4192	0.4259	0.9750	0.2507	0.0217
GpT	0.6044	0.9761	0.4455	0.5772	0.9854	0.6172	0.8891	0.7929	0.0000
GpC	0.8880	0.6506	0.9419	0.7391	0.7011	0.3118	0.3633	0.0000	0.0000
GpG	0.8227	0.9645	0.8405	0.7312	0.2375	0.2618	0.3760	0.0000	0.0000

The following data-tables show a more detailed breakdown of the T-test results for each of the individual sixteen dinucleotides (ApA, ApT, ApC, ApG, TpA, TpT, TpC, TpG, CpA, CpT, CpC and CpG).

ApA

Results for two-tailed T-tests of adjacent repeat masked upstream datasets:
5% level of significance

	Probability	Ho (reject/accept)	datasets - same/different)
UPSTREAM9 / upstream10	0.9599	accept	same
UPSTREAM8 / upstream9	0.7509	accept	same
UPSTREAM7 / upstream8	0.5426	accept	same
UPSTREAM6 / upstream7	0.6380	accept	same
UPSTREAM5 / upstream6	0.2690	accept	same
UPSTREAM4 / upstream5	0.8756	accept	same
UPSTREAM3 / upstream4	0.7838	accept	same
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

ApT

Results for two-tailed T-tests of adjacent repeat masked upstream datasets:
5% level of significance

	Probability	Ho (reject/accept)	datasets - same/different)
UPSTREAM9 / upstream10	0.8356	accept	same
UPSTREAM8 / upstream9	0.7554	accept	same
UPSTREAM7 / upstream8	0.7516	accept	same
UPSTREAM6 / upstream7	0.6211	accept	same
UPSTREAM5 / upstream6	0.6792	accept	same
UPSTREAM4 / upstream5	0.5469	accept	same
UPSTREAM3 / upstream4	0.0259	reject	different
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

ApC

Results for two-tailed T-tests of adjacent repeat masked upstream datasets:
5% level of significance

	Probability	Ho (reject/accept)	datasets - same/different)
UPSTREAM9 / upstream10	0.9350	accept	same

UPSTREAM8 / upstream9	0.6427	accept	same
UPSTREAM7 / upstream8	0.2148	accept	same
UPSTREAM6 / upstream7	0.8497	accept	same
UPSTREAM5 / upstream6	0.5454	accept	same
UPSTREAM4 / upstream5	0.5455	accept	same
UPSTREAM3 / upstream4	0.0400	reject	different
UPSTREAM2 / upstream3	0.3398	accept	same
UPSTREAM1 / upstream2	0.0000	reject	different

ApG

Results for two-tailed T-tests of adjacent repeat masked upstream datasets:
5% level of significance

	Probability	**Ho (reject/accept)**	**datasets - same/different)**
UPSTREAM9 / upstream10	0.3811	accept	same
UPSTREAM8 / upstream9	0.8579	accept	same
UPSTREAM7 / upstream8	0.3419	accept	same
UPSTREAM6 / upstream7	0.9244	accept	same
UPSTREAM5 / upstream6	0.2945	accept	same
UPSTREAM4 / upstream5	0.5660	accept	same
UPSTREAM3 / upstream4	0.3204	accept	same
UPSTREAM2 / upstream3	0.0158	reject	different
UPSTREAM1 / upstream2	0.1135	accept	same

TpA

Results for two-tailed T-tests of adjacent repeat masked upstream datasets:
5% level of significance

	Probability	**Ho (reject/accept)**	**datasets - same/different)**
UPSTREAM9 / upstream10	0.9249	accept	same
UPSTREAM8 / upstream9	0.9395	accept	same
UPSTREAM7 / upstream8	0.6898	accept	same
UPSTREAM6 / upstream7	0.6843	accept	same
UPSTREAM5 / upstream6	0.7342	accept	same
UPSTREAM4 / upstream5	0.5151	accept	same
UPSTREAM3 / upstream4	0.0660	accept	same
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

TpT

Results for two-tailed T-tests of adjacent repeat masked upstream datasets:
5% level of significance

	Probability	**Ho (reject/accept)**	**datasets - same/different)**
UPSTREAM9 / upstream10	0.7247	accept	same
UPSTREAM8 / upstream9	0.5954	accept	same
UPSTREAM7 / upstream8	0.9802	accept	same
UPSTREAM6 / upstream7	0.5337	accept	same
UPSTREAM5 / upstream6	0.5043	accept	same
UPSTREAM4 / upstream5	0.9459	accept	same
UPSTREAM3 / upstream4	0.4899	accept	same
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

TpC

Results for two-tailed T-tests of adjacent repeat masked upstream datasets:
5% level of significance

	Probability	Ho (reject/accept)	datasets - same/different)
UPSTREAM9 / upstream10	0.9036	accept	same
UPSTREAM8 / upstream9	0.9916	accept	same
UPSTREAM7 / upstream8	0.1936	accept	same
UPSTREAM6 / upstream7	0.4273	accept	same
UPSTREAM5 / upstream6	0.9239	accept	same
UPSTREAM4 / upstream5	0.5641	accept	same
UPSTREAM3 / upstream4	0.5864	accept	same
UPSTREAM2 / upstream3	0.9027	accept	same
UPSTREAM1 / upstream2	0.1472	accept	same

TpG

Results for two-tailed T-tests of adjacent repeat masked upstream datasets:
5% level of significance

	Probability	Ho (reject/accept)	datasets - same/different)
UPSTREAM9 / upstream10	0.7158	accept	same
UPSTREAM8 / upstream9	0.7122	accept	same
UPSTREAM7 / upstream8	0.0249	accept	same
UPSTREAM6 / upstream7	0.3846	accept	same
UPSTREAM5 / upstream6	0.8287	accept	same
UPSTREAM4 / upstream5	0.3406	accept	same
UPSTREAM3 / upstream4	0.5468	accept	same
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

CpA

Results for two-tailed T-tests of adjacent repeat masked upstream datasets:
5% level of significance

	Probability	Ho (reject/accept)	datasets - same/different)
UPSTREAM9 / upstream10	0.8333	accept	same
UPSTREAM8 / upstream9	0.4896	accept	same
UPSTREAM7 / upstream8	0.4951	accept	same
UPSTREAM6 / upstream7	0.9889	accept	same
UPSTREAM5 / upstream6	0.7266	accept	same
UPSTREAM4 / upstream5	0.7296	accept	same
UPSTREAM3 / upstream4	0.7634	accept	same
UPSTREAM2 / upstream3	0.0001	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

CpT

Results for two-tailed T-tests of adjacent repeat masked upstream datasets:
5% level of significance

	Probability	Ho (reject/accept)	datasets - same/different)
UPSTREAM9 / upstream10	0.7470	accept	same
UPSTREAM8 / upstream9	0.8033	accept	same
UPSTREAM7 / upstream8	0.4840	accept	same
UPSTREAM6 / upstream7	0.8019	accept	same
UPSTREAM5 / upstream6	0.7387	accept	same
UPSTREAM4 / upstream5	0.7604	accept	same
UPSTREAM3 / upstream4	0.1372	accept	same
UPSTREAM2 / upstream3	0.1344	accept	same
UPSTREAM1 / upstream2	0.0000	reject	different

CpC

Results for two-tailed T-tests of adjacent repeat masked upstream datasets:

5% level of significance			
	Probability	**Ho (reject/accept)**	**datasets - same/different)**
UPSTREAM9 / upstream10	0.9279	accept	same
UPSTREAM8 / upstream9	0.9955	accept	same
UPSTREAM7 / upstream8	0.1172	accept	same
UPSTREAM6 / upstream7	0.4102	accept	same
UPSTREAM5 / upstream6	0.6916	accept	same
UPSTREAM4 / upstream5	0.8679	accept	same
UPSTREAM3 / upstream4	0.4259	accept	same
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

CpG

Results for two-tailed T-tests of adjacent repeat masked upstream datasets:
5% level of significance

	Probability	**Ho (reject/accept)**	**datasets - same/different)**
UPSTREAM9 / upstream10	0.6156	accept	same
UPSTREAM8 / upstream9	0.6119	accept	same
UPSTREAM7 / upstream8	0.5353	accept	same
UPSTREAM6 / upstream7	0.9880	accept	same
UPSTREAM5 / upstream6	0.3980	accept	same
UPSTREAM4 / upstream5	0.7051	accept	same
UPSTREAM3 / upstream4	0.0541	accept	same
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

GpA

Results for two-tailed T-tests of adjacent repeat masked upstream datasets:
5% level of significance

	Probability	**Ho (reject/accept)**	**datasets - same/different)**
UPSTREAM9 / upstream10	0.4921	accept	same
UPSTREAM8 / upstream9	0.6739	accept	same
UPSTREAM7 / upstream8	0.0380	accept	same
UPSTREAM6 / upstream7	0.9813	accept	same
UPSTREAM5 / upstream6	0.4192	accept	same
UPSTREAM4 / upstream5	0.4259	accept	same
UPSTREAM3 / upstream4	0.9750	accept	same
UPSTREAM2 / upstream3	0.2507	accept	same
UPSTREAM1 / upstream2	0.0217	reject	different

GpT

Results for two-tailed T-tests of adjacent repeat masked upstream datasets:
5% level of significance

	Probability	**Ho (reject/accept)**	**datasets - same/different)**
UPSTREAM9 / upstream10	0.6044	accept	same
UPSTREAM8 / upstream9	0.9761	accept	same
UPSTREAM7 / upstream8	0.4455	accept	same
UPSTREAM6 / upstream7	0.5772	accept	same
UPSTREAM5 / upstream6	0.9854	accept	same
UPSTREAM4 / upstream5	0.6172	accept	same
UPSTREAM3 / upstream4	0.8891	accept	same
UPSTREAM2 / upstream3	0.7929	accept	same
UPSTREAM1 / upstream2	0.0000	reject	different

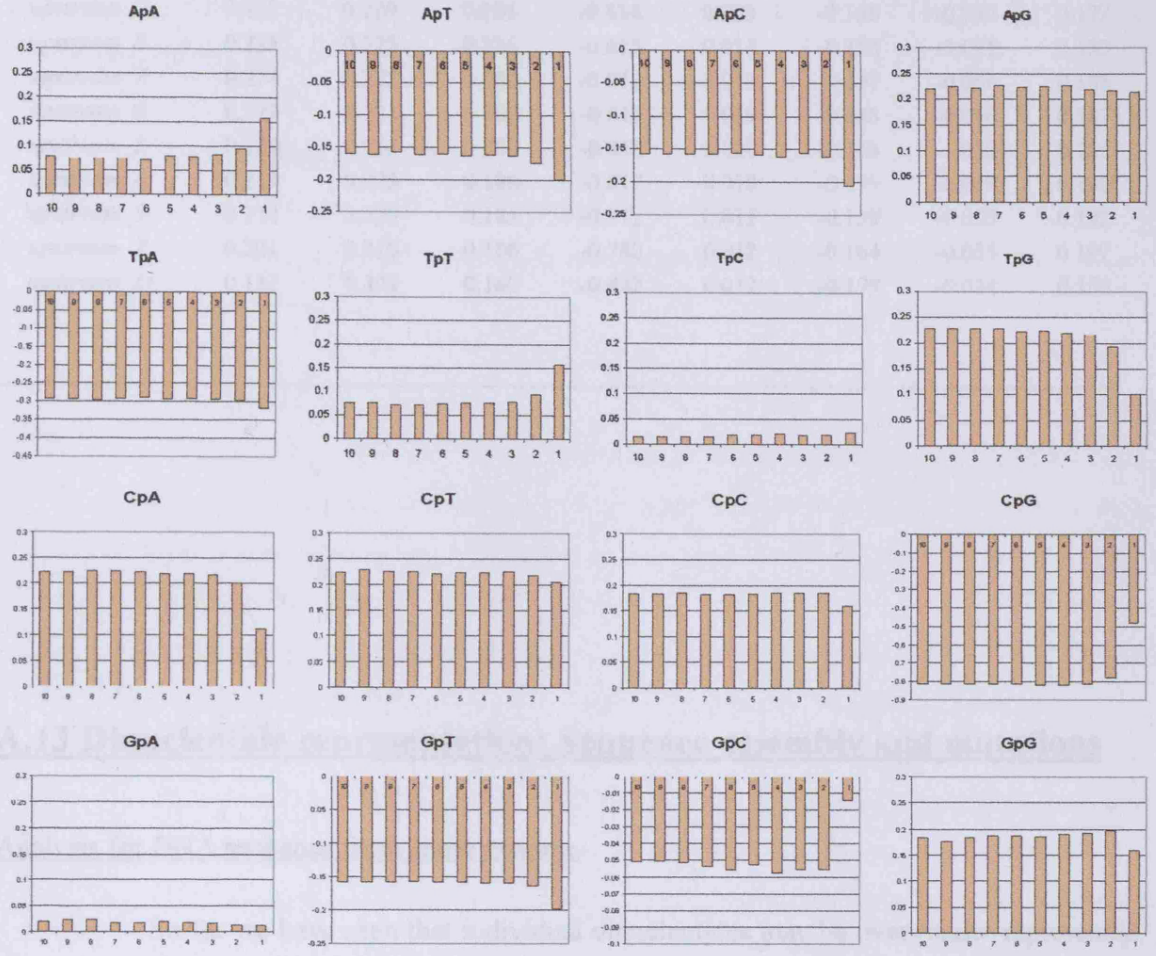
GpC

Results for two-tailed T-tests of adjacent repeat masked upstream datasets: 5% level of significance			
	Probability	**Ho (reject/accept)**	**datasets - same/different)**
UPSTREAM9 / upstream10	0.8880	accept	same
UPSTREAM8 / upstream9	0.6506	accept	same
UPSTREAM7 / upstream8	0.9419	accept	same
UPSTREAM6 / upstream7	0.7391	accept	same
UPSTREAM5 / upstream6	0.7011	accept	same
UPSTREAM4 / upstream5	0.3118	accept	same
UPSTREAM3 / upstream4	0.3633	accept	same
UPSTREAM2 / upstream3	0.0000	reject	different
UPSTREAM1 / upstream2	0.0000	reject	different

GpG			
Results for two-tailed T-tests of adjacent repeat masked upstream datasets: 5% level of significance			
	Probability	**Ho (reject/accept)**	**datasets - same/different)**
UPSTREAM9 / upstream10	0.8227	accept	same
UPSTREAM8 / upstream9	0.9645	accept	same
UPSTREAM7 / upstream8	0.8405	accept	same
UPSTREAM6 / upstream7	0.7312	accept	same
UPSTREAM5 / upstream6	0.2375	Accept	same
UPSTREAM4 / upstream5	0.2618	Accept	same
UPSTREAM3 / upstream4	0.3760	Accept	same
UPSTREAM2 / upstream3	0.0000	Reject	different
UPSTREAM1 / upstream2	0.0000	Reject	different

A.12 Repeats: Dinucleotide representation –descriptive statistics

The following charts and their associated data-tables show the changes in dinucleotide representation (odds ratio: ρxy = fxy/fxfy) across the 10Kb upstream repeat masked sequence (as separate datasets: *upstream1*-to-*upstream10*). The results are shown for each of the sixteen individual dinucleotides.



Dinucleotide distance from randomness (odds ratio-1) for different upstream positional segments

Upstream sequence \ Dinucleotides	ApA	ApT	ApC	ApG	TpA	TpT	TpC	TpG
upstream_10	0.076	-0.160	-0.157	0.220	-0.294	0.076	0.016	0.228
upstream_9	0.074	-0.161	-0.158	0.226	-0.294	0.076	0.016	0.229
upstream_8	0.074	-0.158	-0.158	0.222	-0.296	0.073	0.016	0.228
upstream_7	0.075	-0.161	-0.158	0.226	-0.292	0.073	0.016	0.227
upstream_6	0.072	-0.160	-0.156	0.223	-0.290	0.075	0.019	0.222
upstream_5	0.078	-0.163	-0.159	0.225	-0.293	0.076	0.019	0.223
upstream_4	0.078	-0.162	-0.161	0.226	-0.292	0.077	0.020	0.219
upstream_3	0.083	-0.163	-0.156	0.219	-0.294	0.079	0.017	0.216
upstream_2	0.095	-0.175	-0.162	0.218	-0.296	0.095	0.020	0.193
upstream_1	0.157	-0.199	-0.206	0.213	-0.318	0.157	0.022	0.100

Dinucleotide distance from randomness (odds ratio-1) for different upstream positional segments

Upstream sequence \ Dinucleotides	CpA	CpT	CpC	CpG	GpA	GpT	GpC	GpG
upstream_10	0.223	0.223	0.183	-0.815	0.020	-0.158	-0.051	0.183

upstream_9	0.222	0.229	0.184	-0.814	0.023	-0.158	-0.050	0.177
upstream_8	0.225	0.225	0.186	-0.815	0.023	-0.158	-0.051	0.183
upstream_7	0.224	0.225	0.182	-0.816	0.018	-0.157	-0.054	0.188
upstream_6	0.222	0.221	0.182	-0.815	0.019	-0.158	-0.056	0.187
upstream_5	0.220	0.224	0.183	-0.818	0.020	-0.158	-0.053	0.185
upstream_4	0.219	0.223	0.186	-0.817	0.018	-0.159	-0.058	0.190
upstream_3	0.218	0.225	0.185	-0.812	0.015	-0.159	-0.055	0.192
upstream_2	0.201	0.218	0.186	-0.780	0.012	-0.164	-0.055	0.197
upstream_1	0.112	0.204	0.160	-0.482	0.012	-0.199	-0.014	0.159

A.13 Dinucleotide representation: Sequence assembly and mutations

Analysis for DNA sequence from entire genome

So far we have seen that individual dinucleotides may be over-/under-represented in the upstream sequence and also in other genomic regions. Also, this dinucleotide representation may change across the upstream sequence for the individual dinucleotides.

In this section, based on these results, there will be a discussion as to how the sequence may have come to be the way that it is with respect to dinucleotide representation. The mechanisms via which this process may have taken place will be analysed. This will be done for the genomic DNA in general and then the changes seen across (10Kb) 5' upstream will be considered.

Under-/over-represented dinucleotides: two different potential models

The under-represented dinucleotides were found to be so in all the genomic regions studied. This was true for five out of the possible sixteen, namely; CpG, TpA, ApT, ApC, GpT. Also previous studies have shown that these same dinucleotides are under-represented in genomic DNA of many species. It is true to say that if a set of dinucleotides is under-represented in a sequence, there must also be a set of over-represented dinucleotides.

Theoretically, the sum of proportions of the set of under-represented dinucleotides equals that of the over-represented set.

The dinucleotide is the simplest motif and may be used to study basic sequence properties. A sequence is distant from randomness if its dinucleotides are present at a higher or lower proportion than is expected considering its composition. This situation may come about in two different possible ways;

1. Mutation that is biased:

   This would involve the existence of a 'theoretically random' sequence that then undergoes biased mutation and becomes non-random. Biased mutation means that point mutations, base changes or insertions tend to occur with specific bases in proximity to a specific base.

   In theory under-representation of a dinucleotide may result from a tendency of a base to mutate depending on its neighbouring base. This mutation then results in a new dinucleotide, which is over-represented in the sequence, if it remains stable. Therefore it would be possible, in theory, that the over-represented and under-represented dinucleotides were once all at the random level and that biased point mutations resulted in differential representation which is observed in the sequence today.

   This biased mutation refers to a point mutation that depends on its nearest neighbour. For example, it could be that the nucleotide Y of XpY has a tendency to mutate to Z, yielding new dinucleotide XpZ.

2. Inherent tendency of the DNA to assemble in a particular way:

   The assembly of the DNA double helix may result in a non-random sequence. The formation of dinucleotide steps may favour the pairing of certain nucleotides for structural and energetic reasons. This means that if the conditions were suitable for the formation of DNA steps from their constituent nucleotides, there may be tendency for more XpY pairing to form a step (for example) than X paring with the other nucleotides.

Mutation model for genomic DNA sequence

The reality (existing human DNA) is probably a combination of these two factors conjoined. Specified function (or higher level functionality) of the sequence adds an additional level of complexity and a possible added deviation from the random model. These are the likely reasons for any genomic DNA sequence being non-random with respect to its constituent dinucleotides. I.e., there is probably a basic 'stable' structure onto which is superimposed alternative sequence requirements (and different mutational tendencies) depending on the sequence. The greater the distance from randomness, the higher the occurrence of mutations or the greater the possible bias in sequence assembly.

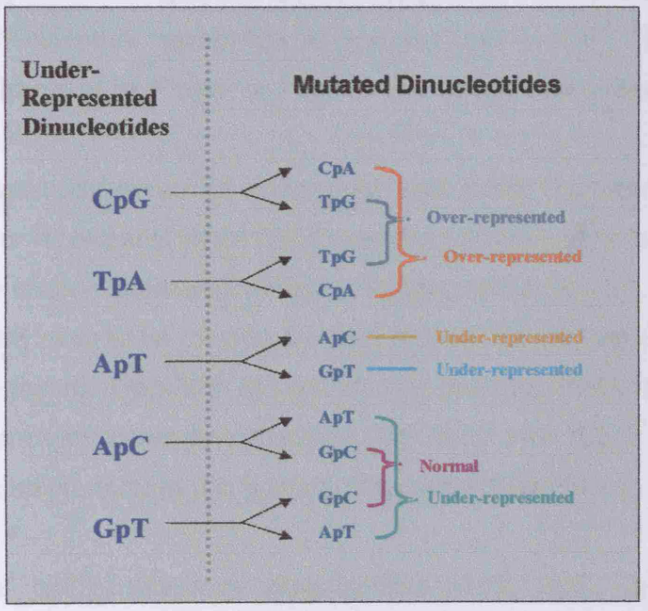


Figure showing under- -represented dinucleotides and their potential transition products in the human genomic DNA:

If a dinucleotide is under-represented in the genomic DNA this may be due to a tendency for one of its bases to mutate. This then results in a new dinucleotide, which is over-represented in the sequence, if it remains stable.

This diagram shows the five under-represented dinucleotides in the genomic sequences that were analyzed and their theoretical mutation products.

The mutated dinucleotides shown are the result of a single base mutation that is either a purine-to-purine or pyrimidine-to-pyrimidine change, i.e. a transition substitution. Therefore each of the under-represented dinucleotides yields two theoretical mutation products, one resulting from a change in its first base, and the other resulting from a change in its second base. The comments next to the mutation products indicate whether in fact those particular dinucleotides are under-represented, over-represented or normal (close to randomness) in the genomic sequences.
The theoretical mutation products of TpA and CpG are over-represented in the genomic sequences. This implies that a transition substitution may lead to the under-representation of TpA and CpG. The under-representation of ApT, ApC and CpT cannot be explained by this type of mutation.

In genomic DNA, the most common type of mutation is thought to be a point mutation or single base substitution. Also, theoretically the most likely substitution is a transition; purine-to-purine change or a pyrimidine-to-pyrimidine change (Zhao et al, 2002). This is thought to be the case because the majority of SNPs are either purine-to-purine or pyrimidine-to-pyrimidine changes. These changes are referred to as 'transitions'. If either of the two bases in a dinucleotide can mutate in this way, two possible dinucleotides result, one resulting from a change of its 5' base, and the other resulting from a change of its 3' base.

If such mutations occur, are wide-spread and stable, the 'new' dinucleotides are then expected to be over-represented in the DNA sequence and the 'old' ones under-represented. For example CpG (an under-represented dinucleotide) may mutate to CpA or TpG. Both CpA and TpG are in actuality seen to be over-represented in the DNA sequences. A similar situation is observed with the mutation products of TpA, also an under-represented dinucleotide. However, of the five under-represented dinucleotides only these two (CpG and TpA) have over-represented mutation products of this type i.e. transition substitutions.

The other under-represented dinucleotides (ApT, ApC, GpT) contain theoretical transition substitution products that in the experimental results of this work are either 'normal' (close to randomness) or are under-represented. The under-representation of these three dinucleotides therefore cannot be explained by a simple and typical transition mutation.

An interesting sequence feature in all the genomic regions studied was the purine and pyrimidine bias of the under- and over-represented dinucleotides. The under-represented dinucleotides all contained one purine and one pyrimidine base. In contrast to this the over-represented dinucleotides (except for TpG and CpA that are the theoretical mutation products of CpG and TpA) were composed of either two purine or two pyrimidine bases. Therefore the

under-representation of ApT, ApC and GpT cannot be explained in terms of simple transition mutations. The mechanism via which this may take place (i.e. the mechanism via which this type of sequence has been generated) is unknown.

It is therefore impossible from these results to explain that these sequences arose solely as a result of point mutations that are 'transitions'. This is the case even if the mutations are non-biased i.e. random. Even if mutations were not biased in that they may not depend on their nearest neighbour, purines would still be required to mutate to purines and pyrimidines to pyrimidines for this type of transition mutation model.

Therefore it may be that an alternative form of mutation generated the present day sequence. It could be that transition (purine to pyrimidine point mutations) or vice versa have resulted in this observed over- and under-representation of dinucleotides. Since these are far less common than transitions, this would also seem to be an unlikely explanation for the observed results.

What does SNP data show?

In a study whereby approximately 2.5 million SNPS were investigated it was found that the relative proportions of transition substitutions was 65.58% and transversions; 34.42% (Zhao et al, 2002). This means that the transition rate would be much higher than transversion, although there is an added complication. Transition and transversion occurrence is biased by the nearest neighbouring nucleotide (Morton, 1995, Morton et al, 2005, Zhang et al, 2004). For example, it has been shown that in rice and maize chloroplast non-coding DNA, the type of substitution is dependent upon the A+T verses C+G content of neighbouring bases (Morton, 1995). When both the 5' and 3' flanking are either C or G, 75% of substitutions are transitions and 25% are transversions. When both 5' and 3' flanking bases are either A or T, 43% of substitutions are transitions and 57% are transversions. Therefore we see that transversions whilst altogether less common than transitions, could in theory occur at a higher rate in locations that are flanked by two weak residues.

This means that a mutation model for the DNA sequence based on these base substitutions is complex. It is known from the dinucleotide representation results that all under-represented dinucleotides are composed of one purine and one pyrimidine base. In order to explain this representation effect in terms of mutation, this result must be attributed (at least in part) to transversion substitutions. In addition, the occurrence of transversions would probably have to be in excess of transitions, as will be explained.

The balance of theoretical transitions verses transversions

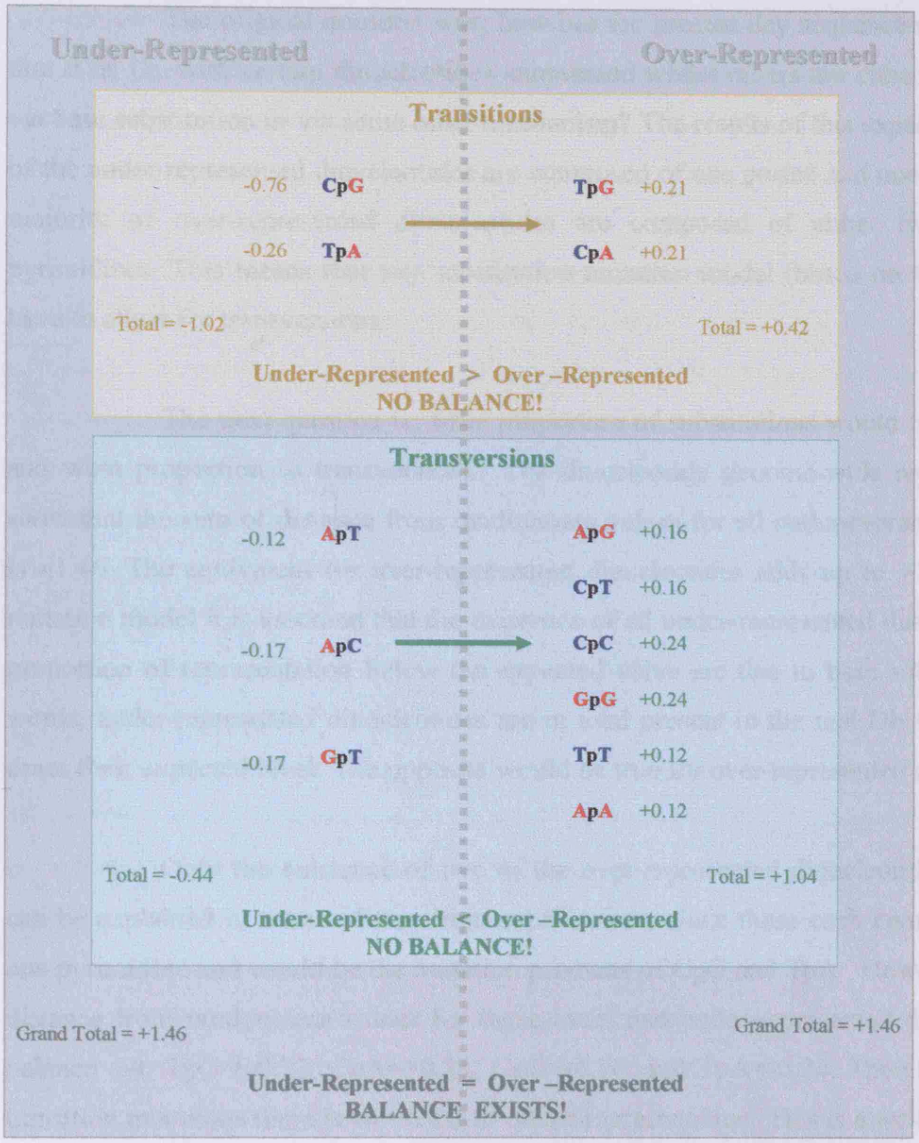


Diagram showing a balance sheet of substitution mutations as a way explaining the existence of over-represented and under-represented dinucleotides within the *whole genome* sequence.

The left side of the balance sheet shows the under-represented dinucleotides and on the right side are the over-represented dinucleotides. Purine bases are shown in red and pyrimidines in blue.

Theoretical mutation products of the under-represented dinucleotides are shown with their representation values. These single base substitutions may be in the form of either transitions or transversions.

The balance sheet shows that for the possible transition events, the proportion of under-representation is higher in total than that for over-representation. In contrast, for

transversions the opposite is true. The proportion of over-representation is higher than that for under-representation.

The original question was; how has the present day sequence come to be the way that it is, i.e. with certain dinucleotides suppressed whilst others are enhanced; did this occur via base substitution or via some other mechanism? The results of this experiment show that all of the under-represented dinucleotides are composed of one purine and one pyrimidine and the majority of over-represented dinucleotides are composed of either two purines or two pyrimidines. This means that any substitution mutation model (based on these results) would have to allow for transversions.

The next question is; what proportion of substitutions would be due to transitions and what proportion to transversions? The dinucleotide genome-wide representation results show that the sum of distance from randomness values for all under-represented dinucleotides is −1.46. The equivalent for over-represented dinucleotides adds up to +1.46. Based on this mutation model it is assumed that the existence of all under-represented dinucleotides and their proportion of representation below the expected value are due to base substitutions. In other words, under-represented dinucleotides are in total present in the real DNA sequence at -1.46 times their expected level. The opposite would be true for over-represented dinucleotides.

Only the existence of two of the over-represented dinucleotides (TpG and CpA) can be explained in terms of transition substitutions, since these each contain one purine and one pyrimidine and would be the mutation products of CpG and TpA . However, the sum of the distance from randomness values for these over- and under-represented dinucleotides do not balance out. TpG=+0.21, CpA=+0.21, CpG=-0.76, and TpA=-0.26. Therefore for the sum of transition mutations there is an excess of under-representation. This is a well-known problem in genomic DNA across the species.

All of the remaining over-represented dinucleotides (ApG, CpT, CpC, GpG, TpT, ApA) are composed of either two purines or two pyrimidines and therefore could only have arisen as a result of transversions. However, within this category too, there is a lack of balance in over- and under-representation values. The representation values for these over-represented dinucleotides in total is =+1.04, and for the under-represented dinucleotides the total is, -0.44. Therefore for these transversion mutations there seems to be an excess of over-representation in terms of the magnitude of the representation values.

Since in total (for both transitions and transversions) there is an equal balance for over- and under-represented dinucleotides, this lack of balance for the transversions and transitions individually could possibly be attributed to transversion products for CpG and TpA.

In other words, perhaps these dinucleotides undergo transversion substitutions and not just transitions. Either way, dinucleotides that would be the products of transversion substitutions possess a total distance from randomness ratio in excess of +1.04 whereas the equivalent for transitions is +0.42. This implies that the occurrence of transversions may be more than twice as high as transitions in the human genomic DNA since these are values taken from the entire genome.

It is known from SNP data (outlined above) that in general transitions constitute 65.58% of substitutions whilst transversions constitute 34.42%. So do the relative proportions of transitions verses transversions estimated from this representation experiments fit in with those from the SNP data? The answer to this question is no. The SNP results show that overall transitions are almost twice as likely as transversions. The representation data in this project implies that transversions are more than twice as likely as transitions. The two observations would therefore seem contradictory.

The SNP data though is more complicated than the overall transition/transversion percentages given above. For example, SNP data shows that there as an increased occurrence of transversion substitutions if both flanking bases are either A or T to 57%. It seems unlikely though that transversions could account for the relative representation values of the dinucleotides. This draws on the likelihood for the existence of other mechanisms that have brought about the present-day genomic DNA sequence.

Is the present dinucleotide representation of the sequence due to substitution mutations or something else? When the discrepancies between the SNP data and the representation of dinucleotides are considered, it would seem that the likely occurrence of transversions (according to the SNP data) may not be sufficient to explain the representation of dinucleotides in the genomic DNA. Therefore it is possible that other mechanisms are responsible.

It may be for example, that biased single base insertions have occurred in the DNA sequence specifically in sites that yield a greater number of purine only or pyrimidine only dinucleotides. However, this also seems unlikely. Another alternative is that the over-represented dinucleotides at least in part have arisen not from the mutations described above, but from direct repeats or repeat expansions that tend to include (for some unknown reason) either purine only dinucleotides or pyrimidine only dinucleotides. This would possibly explain their over-representation arising without mutation.

An alternative explanation is that perhaps the observed representation of dinucleotides is due to a tendency for the DNA to assemble in a particular way and is not entirely due to mutations. This means that perhaps DNA sequence had a tendency for assembly in such a way that suppressed steps with a purine neighbouring a pyrimidine and enhanced the assembly of steps with wither two purines or two pyrimidines.

This type of tendency for base steps may be essential to the most basic structure of the DNA helix and its stability. Any variation within genomic DNA therefore is likely to still harbour this basic property, the variation itself altering the structure but only to a certain degree. This would provide a possible explanation as to why the representation of dinucleotides cannot be entirely accounted for by the most likely mutation events that have been described above. The SNP data may be more indicative of mutational tendencies of the DNA than the dinucleotide representation data.

Analysis for DNA sequence across the upstream

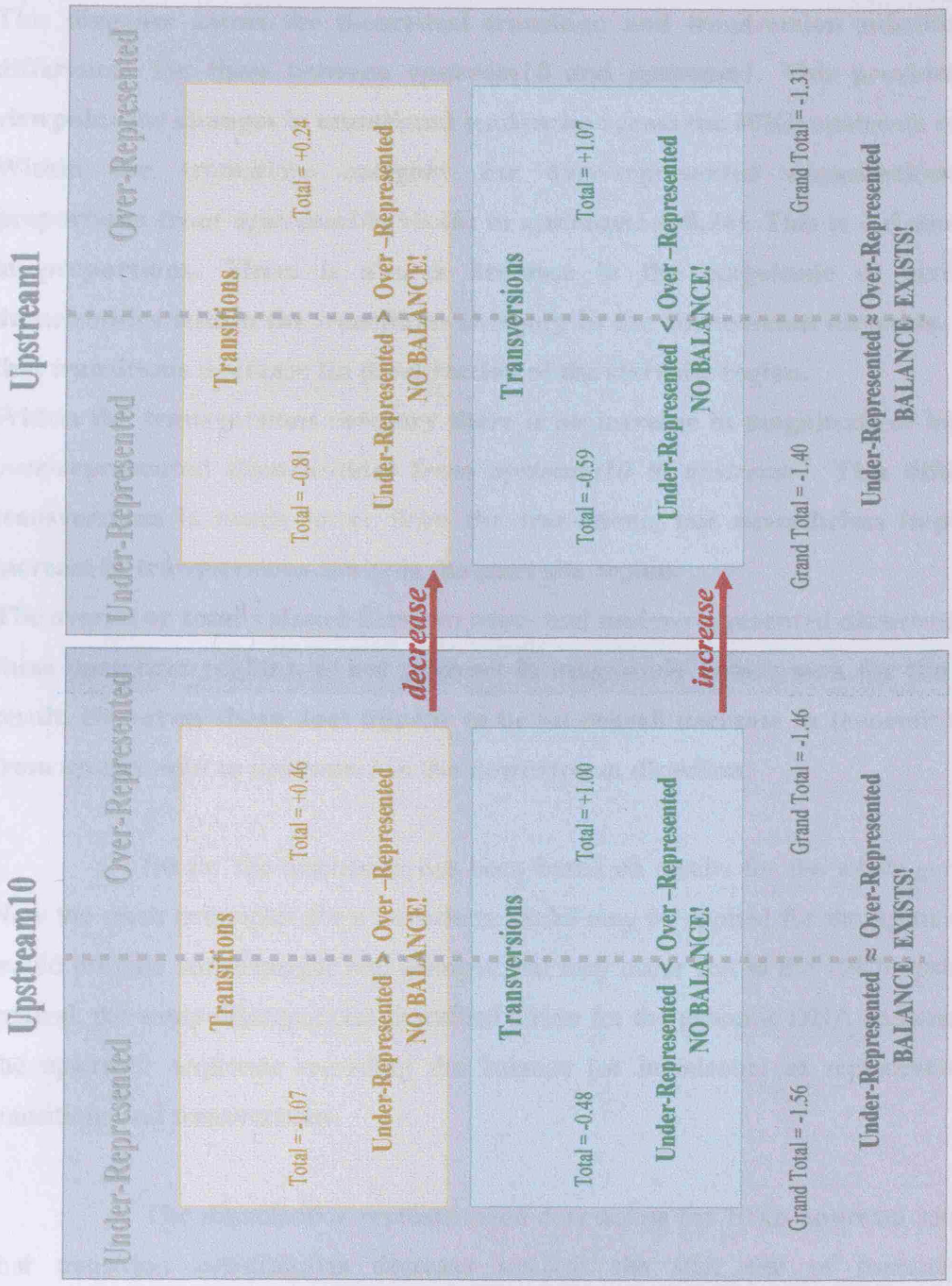


Diagram showing a summary balance sheet of substitution mutations within the *upstream10* and *uptsream1* regions. This is presented a way explaining the existence of over-represented and under-represented dinucleotides. (The balance sheet is similar to the avoe figure except that the representation values of the individual dinucleotides is not shown)

The balance sheet shows that for the possible transition events, the proportion of under-representation is higher in total than that for over-representation. In contrast, for transversions the opposite is true. The proportion of over-representation is higher than that for under-representation. Therefore this imbalance is similar that seen for the *whole genome*.

This diagram shows the theoretical transition and transversion substitutions and the differences for these between *upstream10* and *upstream1*. This provides one possible viewpoint for changes in mutational tendencies across the 10Kb upstream region.
Within the transitions category the over-represented dinucleotides decrease is proportions from *upstream10* (+0.46) to *upstream1* (+0.24). This is a dramatic difference in proportions. There is also a decrease in the magnitude of under-represented dinucleotides within the transitions category in the downstream direction. This suggests that transitions decrease tin the direction of the start site region.
Within the transversions category there is an increase in magnitude of both under- and over-represented dinucleotides from *upstream10* to *upstream1*. This difference for the transversions is much lower than the transitions, but nevertheless implies a possible increase in transversions towards the start site region.
The overall or total balance between over- and under-represented dinucleotides in both of these upstream regions, is not as exact in magnitude as was seen for the *whole genome* result. However, there does appear to be an overall decrease in theoretical substitutions from *upstream10* to *upstream1* in the downstream direction.

So far the discussion has been based on results for the *whole genome* sequence. Now the same principles for a mutations model may be applied for the upstream region. This would provide some insight into changes that may occur across the 10Kb upstream region. In general, the same discrepancies described above for the genomic DNA sequence also apply to the upstream sequence regarding the balance (or imbalance) of representation values for transitions and transversions.

The dinucleotide representation data across the 10Kb upstream sequences suggest that transition substitutions decrease towards the start site of transcription, whereas transversions increase. However, this same data also implies that in total the level of these theoretical substitutions decrease in the downstream direction, towards the start site.

According to SNP data analysis (Guo et al, 2005), the total number of substitutions increases in the direction of the start site. Also, according to these results, towards the start site a higher proportion of SNPs are transversions in comparison with the more upstream sequence.

This is in agreement with the changes in dinucleotide representation seen in this experiment and their implications for transitions and transversions across the upstream sequence. It appears therefore that the R/Y sequence becomes less conserved towards the TSS. The reason for this is unknown.

Appendix B

B.1 Distance from randomness: the effect of repeats on the ATCG upstream sequence

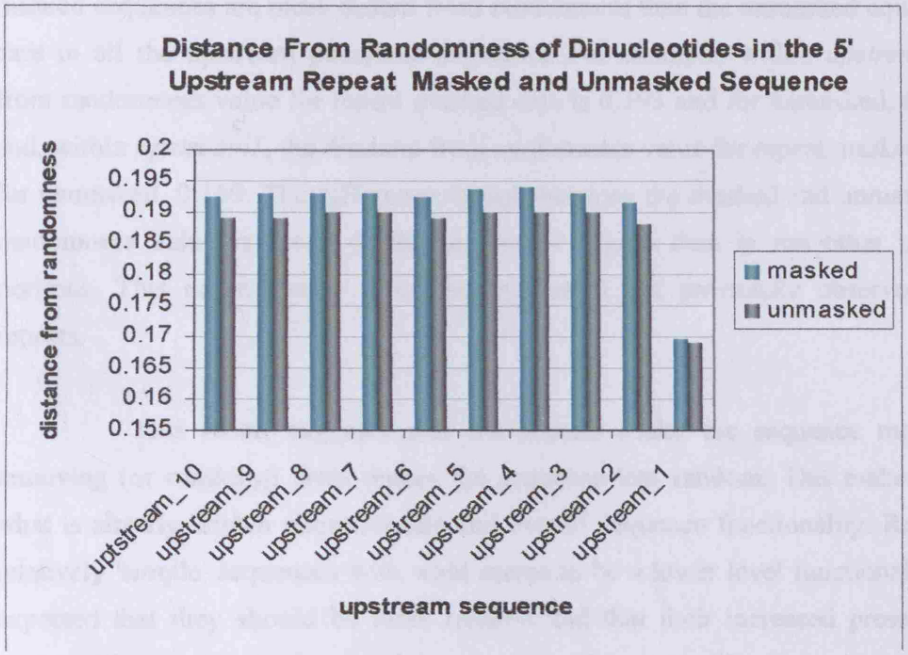


Table of distance from randomness values across the upstream repeat masked sequence

Upstream region	distance from randomness
upstream_10	0.193
upstream_9	0.193
upstream_8	0.193
upstream_7	0.193
upstream_6	0.192
upstream_5	0.194
upstream_4	0.194
upstream_3	0.193
upstream_2	0.192
upstream_1	0.170

This graph shows the distance from randomness values of the different upstream positional segments both for repeat masked and unmasked sequence datasets. The median value is given for each dataset of sequences. The general trend is that the sequence becomes closer to randomness in the downstream direction for both the repeat masked and unmasked datasets.

The difference between the masked and unmasked sequences is that the repeat masked or repeat-free sequences (in all the upstream positional segments) are more distant from randomness than the unmasked equivalent. It is expected that the elimination of repeats should make the sequences less random.

Within the 10Kb upstream, the sequence becomes closer to randomness in the downstream direction (towards *upstream1*) for both the repeat masked and unmasked datasets. This is the general trend for the unmasked sequence and it remains unchanged for repeat masked sequences.

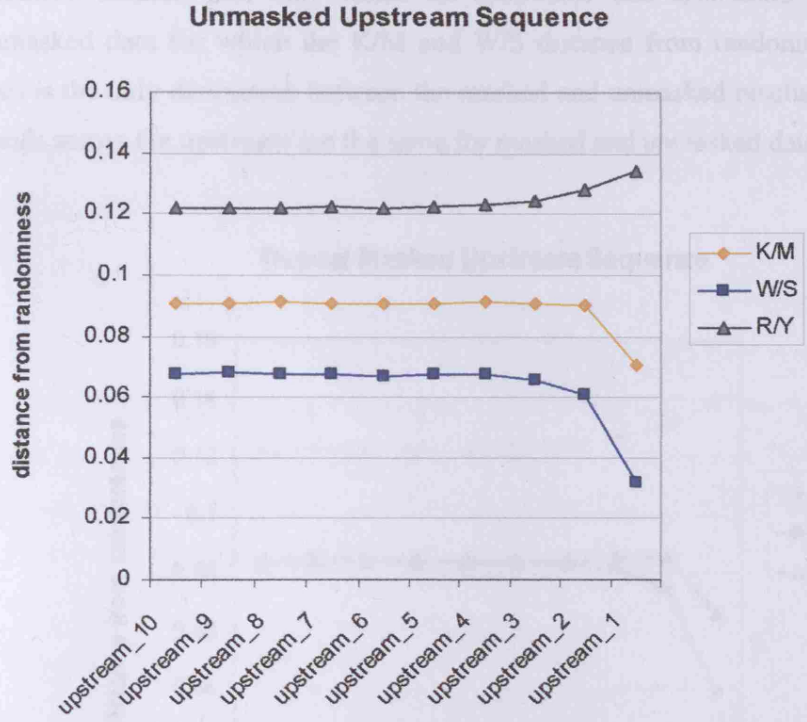
The difference between the masked and unmasked sequences is that the repeat masked sequences are more distant from randomness than the unmasked equivalent. This is the case in all the upstream positional segments. For example, within *upstream10*, the distance from randomness value for repeat masked data is 0.193 and for unmasked, 0.189. At the other end, within *upstream1*, the distance from randomness value for repeat masked data is 0.170 and for unmasked, 0.169. The difference though between the masked and unmasked distance from randomness value is lower in the *upstream1* region than in the other upstream sequence portions. This makes sense since the upstream1 (as previously observed) contains fewer repeats.

This result suggests that the repeats make the sequence more random, since removing (or masking) them makes the sequence less random. This makes sense in light of what is already known about repeats and overall sequence functionality. Repeats are it seems relatively 'simple' sequences with what seems to be a lower level functionality. Therefore it is expected that they should be more random and that their increased presence should confer random-like characteristics on the sequence. See appendix for statistical analyses on the difference between real and random datasets across the 10Kb upstream sequence as was done for the unmasked datasets.

B.2 Distance from randomness: The effect of repeats on the R/Y, W/S and K/M upstream sequence

Distance from randomness values are given for the R/Y, W/S and K/M sequence translations of the original ATCG sequence. The K/M sequence refers to a conversation of A and C to X, and T and G to Y. K/M is used here as a non- R/Y and a non- W/S control. Distance from randomness results (average relative abundance dinucleotide profiles) show that all three types are generally non-random. However, identical sequences possesses different relative distance from randomness values depending on the translation (i.e. K/M, W/S and R/Y).

The results suggest a higher order association of purines and pyrimidines in the formation of dinucleotides. This in turn shows an importance of the R/Y sequence. The weak and strong bases do not show this same level of association. The K/M control lies between the two in this respect. Therefore there is a greater level of R/Y sequence importance in the 10Kb upstream, which increases in the TSS direction.



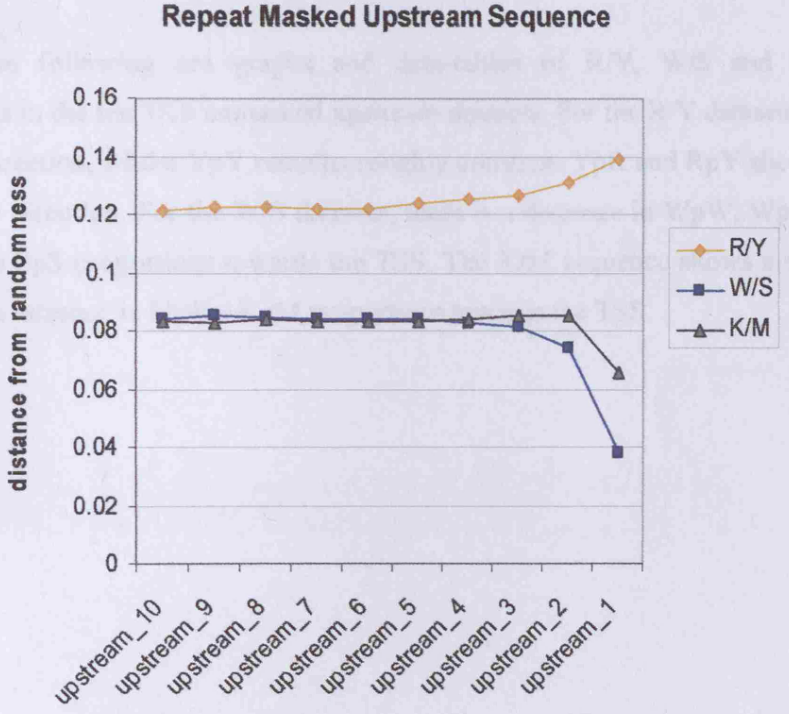
Distance from randomness- Unmasked datasets			
	K/M	W/S	R/Y
upstream_10	0.091	0.067	0.123
upstream_9	0.091	0.066	0.124
upstream_8	0.090	0.061	0.128
upstream_7	0.070	0.032	0.134
upstream_6	0.000	0.000	0.000
upstream_5	0.000	0.000	0.000
upstream_4	0.000	0.000	0.000
upstream_3	0.091	0.067	0.123
upstream_2	0.091	0.066	0.124
upstream_1	0.090	0.061	0.128

Graph and data-table of distance from randomness for K/M, W/S and R/Y unmasked upstream sequences. A value of zero is expected for a random sequence.

The W/S and K/M sequence becomes increasingly closer to the random expectation between *upstream5* and *upstream1*, in the TSS direction. The opposite is true for R/Y where the sequence becomes less random towards the TSS.

Also, throughout the 10Kb upstream the R/Y sequence is the least random, followed by the K/M sequence and then the W/S.

The distance from randomness trends across the upstream are the same for unmasked and masked sequences. For repeat masked sequences, the R/Y distance from randomness profile is similar as for the equivalent unmasked upstream sequence. For the repeat masked upstream, the K/M and W/S distance from randomness profile is almost identical across the upstream datasets and only varies for *upstream1* and *upstream2*. This is different to the unmasked data for which the K/M and W/S distance from randomness values are different. This is the only distinction between the masked and unmasked results. Despite this, the overall trends across the upstream are the same for masked and unmasked datasets.



Distance from randomness Repeat masked datasets	K/M	W/S	R/Y
upstream_10	0.125	0.082	0.084
upstream_9	0.126	0.082	0.085
upstream_8	0.131	0.074	0.085
upstream_7	0.139	0.038	0.066
upstream_6	0.000	0.000	0.000
upstream_5	0.000	0.000	0.000
upstream_4	0.000	0.000	0.000
upstream_3	0.125	0.082	0.084
upstream_2	0.126	0.082	0.085
upstream_1	0.131	0.074	0.085

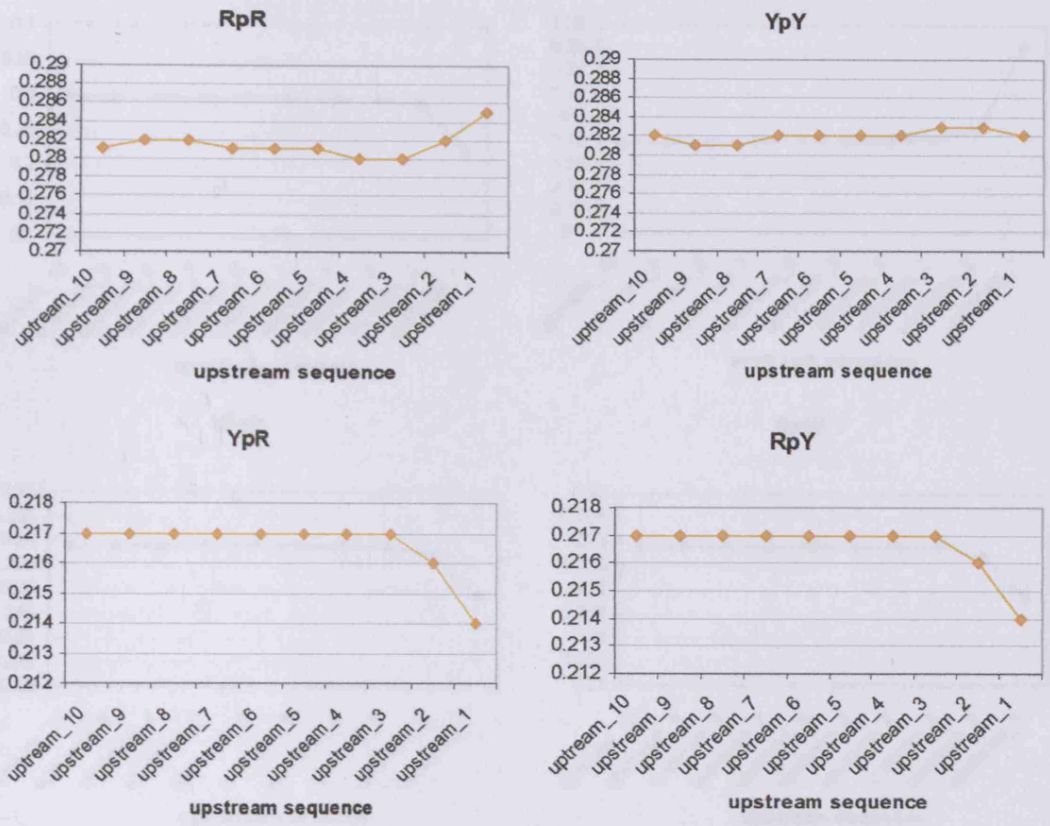
Graph and data-table of distance from randomness (average relative abundance dinucleotide) for K/M, W/S and R/Y repeat masked upstream sequences. A value of zero is expected for a random sequence.

The W/S and K/M sequence becomes increasingly closer to the random expectation between *upstream5* and *upstream1*, in the TSS direction. The K/M and W/S distance from randomness profile is almost identical across the upstream datasets and only varies for *upstream1* and *upstream2*. For R/Y the sequence becomes less random towards the TSS.

B3: Individual dinucleotide proportions: R/Y, W/S and K/M unmasked upstream sequences

The following are graphs and data-tables of R/Y, W/S and K/M dinucleotide proportions in the ten 1Kb unmasked upstream datasets. For the R/Y datasets, RpR increases in the TSS direction, whilst YpY remains roughly constant. YpR and RpY show decreased levels in the TSS direction. For the W/S datasets, there is a decrease in WpW, WpS and SpW and an increase in SpS proportions towards the TSS. The K/M sequence shows a decrease in MpM / KpK and a increase in MpK / KpM proportions towards the TSS.

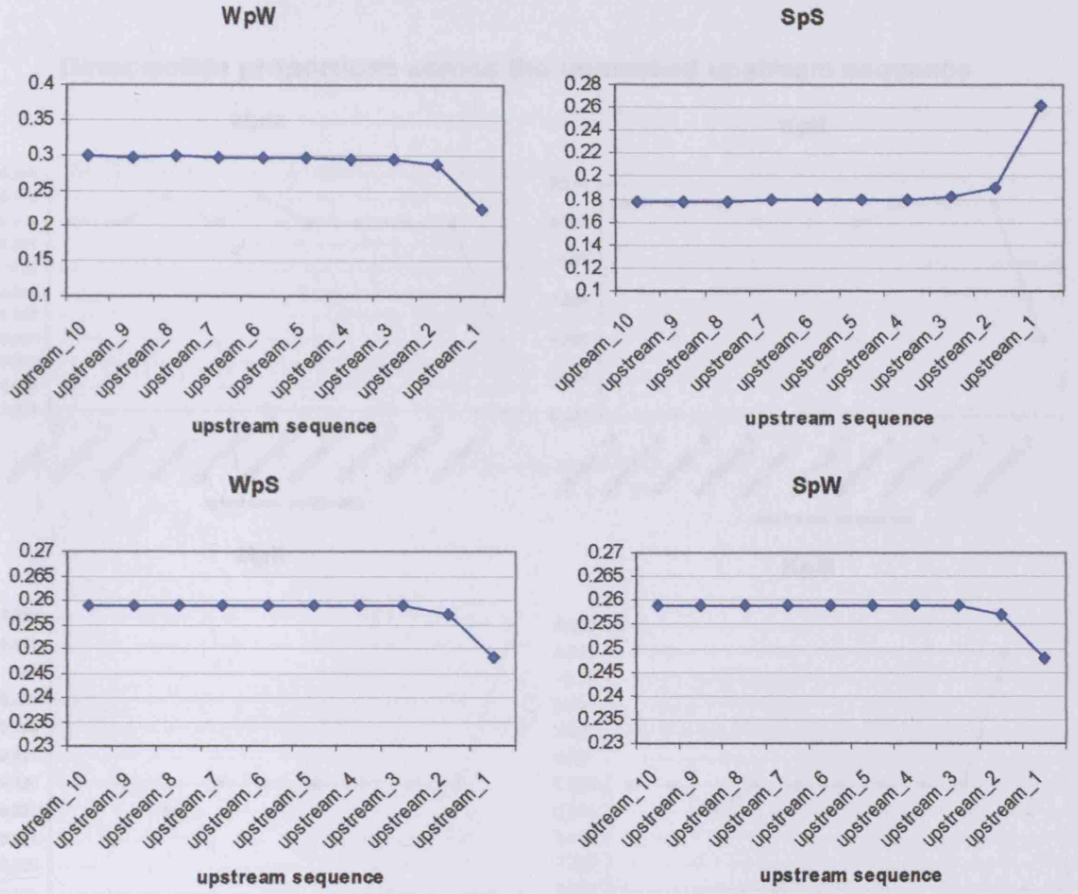
Dinucleotide proportions across the upstream sequence



Unmasked upstream datasets
Individual DINUCLEOTIDES: proportions (median)

	RpR	RpY	YpR	YpY
upstream_10	0.281	0.217	0.217	0.282
upstream_9	0.282	0.217	0.217	0.281
upstream_8	0.282	0.217	0.217	0.281
upstream_7	0.281	0.217	0.217	0.282
upstream_6	0.281	0.217	0.217	0.282
upstream_5	0.281	0.217	0.217	0.282
upstream_4	0.28	0.217	0.217	0.282
upstream_3	0.28	0.217	0.217	0.283
upstream_2	0.282	0.216	0.216	0.283
upstream_1	0.285	0.214	0.214	0.282

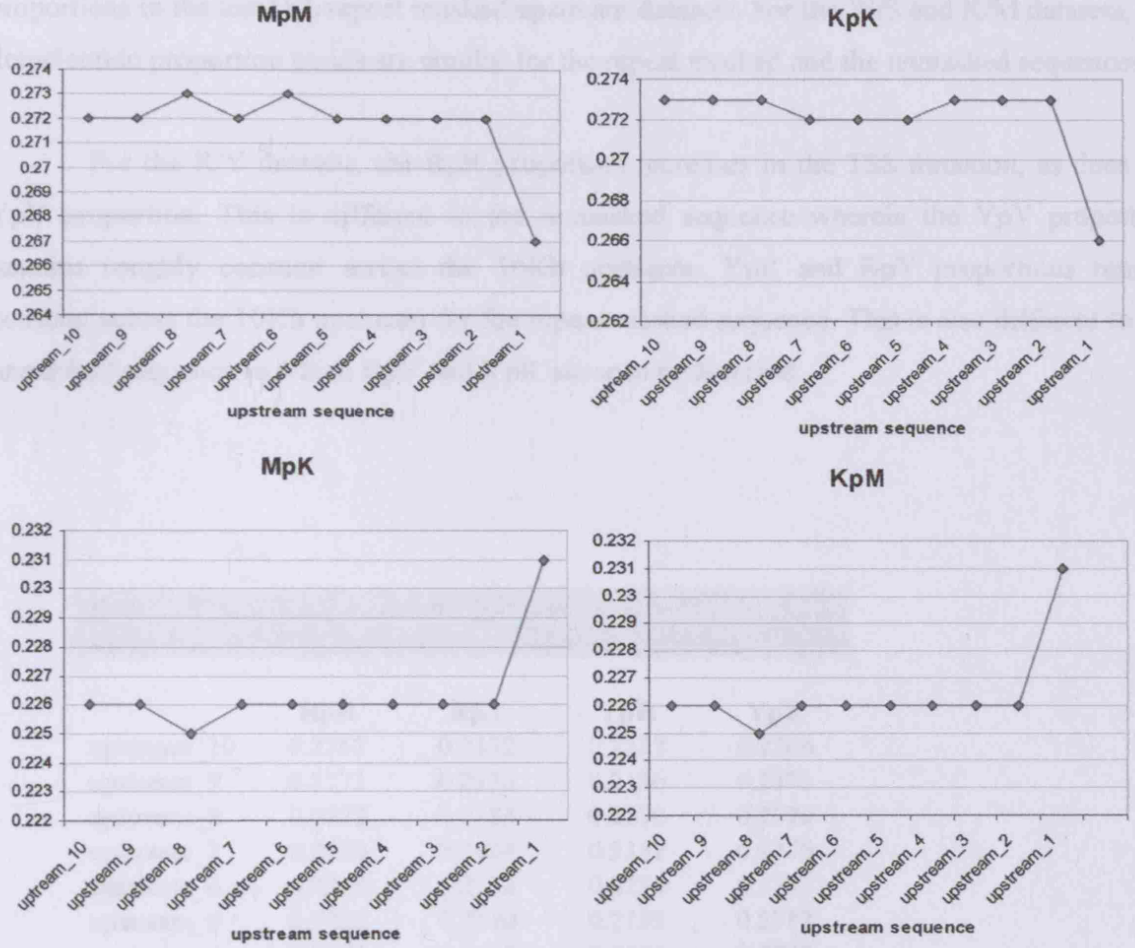
Dinucleotide proportions across the upstream sequence



Unmasked upstream datasets
Individual DINUCLEOTIDES: proportions (median)

	WpW	WpS	SpW	SpS
upstream_10	0.298	0.259	0.259	0.178
upstream_9	0.297	0.259	0.259	0.178
upstream_8	0.298	0.259	0.259	0.178
upstream_7	0.296	0.259	0.259	0.179
upstream_6	0.297	0.259	0.259	0.179
upstream_5	0.297	0.259	0.259	0.179
upstream_4	0.294	0.259	0.259	0.180
upstream_3	0.293	0.259	0.259	0.183
upstream_2	0.286	0.257	0.257	0.190
upstream_1	0.224	0.248	0.248	0.262

Dinucleotide proportions across the unmasked upstream sequence



Unmasked upstream datasets
Individual DINUCLEOTIDES: proportions (median)

	MpM	MpK	KpM	KpK
upstream_10	0.272	0.226	0.226	0.273
upstream_9	0.272	0.226	0.226	0.273
upstream_8	0.273	0.225	0.225	0.273
upstream_7	0.272	0.226	0.226	0.272
upstream_6	0.273	0.226	0.226	0.272
upstream_5	0.272	0.226	0.226	0.272
upstream_4	0.272	0.226	0.226	0.273
upstream_3	0.272	0.226	0.226	0.273
upstream_2	0.272	0.226	0.226	0.273
upstream_1	0.267	0.231	0.231	0.266

B4: Individual dinucleotide proportions: R/Y, W/S and K/M repeat masked upstream sequences

The following are graphs and data-tables of R/Y, W/S and K/M dinucleotide proportions in the ten 1Kb repeat masked upstream datasets. For the W/S and K/M datasets, the dinucleotide proportion trends are similar for the repeat masked and the unmasked sequences.

For the R/Y datasets, the RpR proportion increases in the TSS direction, as does the YpY proportion. This is different to the unmasked sequence wherein the YpY proportion remains roughly constant across the 10Kb upstream. YpR and RpY proportions remain constant across the 10Kb upstream for the repeat masked sequence. This is also different to the unmasked sequence in which RpY and YpR are seen to decrease.

REPEAT MASKED upstream datasets
Individual DINUCLEOTIDES: proportions (median)

	RpR	RpY	YpR	YpY
upstream_10	0.2762	0.2132	0.2137	0.2766
upstream_9	0.2773	0.2133	0.2136	0.2773
upstream_8	0.2772	0.2134	0.2138	0.2770
upstream_7	0.2770	0.2134	0.2137	0.2776
upstream_6	0.2760	0.2132	0.2136	0.2787
upstream_5	0.2773	0.2130	0.2135	0.2787
upstream_4	0.2779	0.2132	0.2136	0.2789
upstream_3	0.2778	0.2132	0.2136	0.2801
upstream_2	0.2805	0.2128	0.2130	0.2812
upstream_1	0.2840	0.2126	0.2127	0.2830

REPEAT MASKED upstream datasets
Individual DINUCLEOTIDES: proportions (median)

	WpW	WpS	SpW	SpS
upstream_10	0.3108	0.2539	0.2540	0.1472
upstream_9	0.3102	0.2546	0.2545	0.1486
upstream_8	0.3117	0.2546	0.2545	0.1477
upstream_7	0.3089	0.2544	0.2545	0.1504
upstream_6	0.3092	0.2545	0.2547	0.1497
upstream_5	0.3133	0.2542	0.2545	0.1481
upstream_4	0.3139	0.2546	0.2547	0.1505
upstream_3	0.3107	0.2550	0.2550	0.1527
upstream_2	0.3010	0.2544	0.2545	0.1644
upstream_1	0.2105	0.2466	0.2465	0.2731

REPEAT MASKED upstream datasets
Individual DINUCLEOTIDES: proportions (median)

	MpM	MpK	KpM	KpK
upstream_10	0.2665	0.2249	0.2250	0.2684
upstream_9	0.2669	0.2250	0.2250	0.2684
upstream_8	0.2668	0.2247	0.2248	0.2685
upstream_7	0.2673	0.2249	0.2249	0.2677
upstream_6	0.2670	0.2250	0.2251	0.2685
upstream_5	0.2675	0.2251	0.2252	0.2678
upstream_4	0.2679	0.2250	0.2251	0.2684
upstream_3	0.2685	0.2251	0.2251	0.2690
upstream_2	0.2691	0.2257	0.2257	0.2698
upstream_1	0.2656	0.2316	0.2316	0.2647

B5: Individual dinucleotide representations: R/Y, W/S and K/M unmasked upstream sequences

The following are data-tables of representation values for all possible individual dinucleotides in the R/Y, W/S, and K/M unmasked sequences. All the R/Y dinucleotides become more distant from the random sequence expectation (zero value) in the TSS direction. In contrast all W/S and K/M dinucleotides become closer to the random expectation in the same direction.

Unmasked upstream datasets
Individual DINUCLEOTIDES: (ODDS Ratio-1)

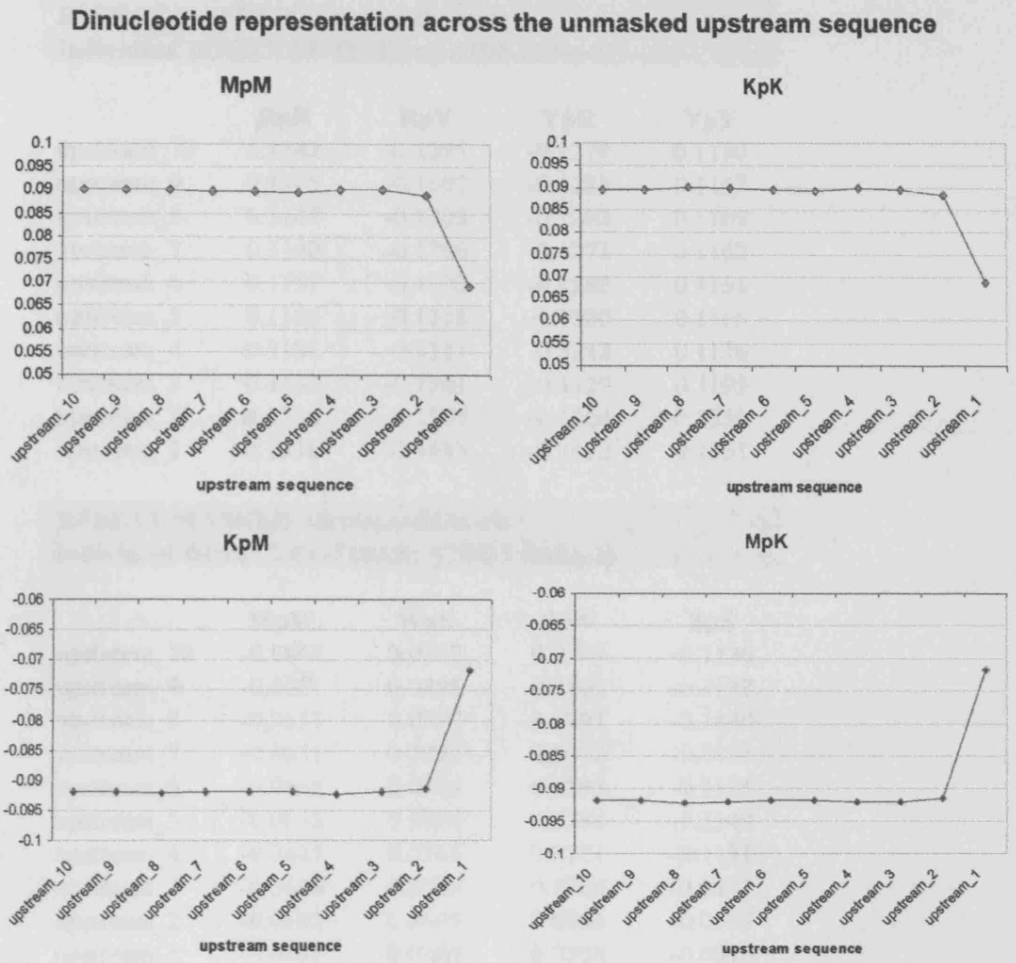
	RpR	RpY	YpR	YpY
upstream_10	0.1200	-0.1234	-0.1234	0.1204
upstream_9	0.1201	-0.1236	-0.1235	0.1207
upstream_8	0.1201	-0.1235	-0.1235	0.1198
upstream_7	0.1205	-0.1239	-0.1239	0.1209
upstream_6	0.1209	-0.1238	-0.1237	0.1200
upstream_5	0.1210	-0.1244	-0.1244	0.1207
upstream_4	0.1219	-0.1246	-0.1248	0.1210
upstream_3	0.1238	-0.1260	-0.1260	0.1218
upstream_2	0.1267	-0.1303	-0.1303	0.1258
upstream_1	0.1320	-0.1359	-0.1360	0.1335

Unmasked upstream datasets
Individual DINUCLEOTIDES: (ODDS Ratio-1)

	WpW	WpS	SpW	SpS
upstream_10	-0.0503	0.0666	0.0666	-0.0856
upstream_9	-0.0510	0.0672	0.0673	-0.0858
upstream_8	-0.0502	0.0670	0.0668	-0.0863
upstream_7	-0.0508	0.0666	0.0668	-0.0856
upstream_6	-0.0506	0.0661	0.0662	-0.0847
upstream_5	-0.0504	0.0666	0.0665	-0.0853
upstream_4	-0.0509	0.0666	0.0666	-0.0855
upstream_3	-0.0502	0.0651	0.0650	-0.0829
upstream_2	-0.0473	0.0602	0.0601	-0.0754
upstream_1	-0.0318	0.0330	0.0323	-0.0307

Unmasked upstream datasets
Individual DINUCLEOTIDES: (ODDS Ratio-1)

	MpM	MpK	KpM	KpK
upstream_10	0.0900	-0.0919	-0.0919	0.0895
upstream_9	0.0901	-0.0919	-0.0919	0.0896
upstream_8	0.0901	-0.0921	-0.0923	0.0902
upstream_7	0.0898	-0.0920	-0.0920	0.0900
upstream_6	0.0897	-0.0919	-0.0919	0.0897
upstream_5	0.0896	-0.0919	-0.0918	0.0893
upstream_4	0.0901	-0.0923	-0.0920	0.0900
upstream_3	0.0902	-0.0920	-0.0920	0.0896
upstream_2	0.0889	-0.0915	-0.0915	0.0882
upstream_1	0.0690	-0.0718	-0.0718	0.0688



B6: Individual dinucleotide representations: R/Y, W/S and K/M repeat masked upstream sequences

The following are data-tables of representation values for all possible individual dinucleotides in the R/Y, W/S, and K/M repeat masked sequences. All the R/Y dinucleotides become more distant from the random sequence expectation (zero value) in the TSS direction. In contrast all W/S and K/M dinucleotides become closer to the random expectation in the same direction. Therefore repeat masked sequence show similar trends across the 10Kb sequence to unmasked sequences with respect to the dinucleotide representation.

REPEAT MASKED upstream datasets
Individual DINUCLEOTIDES: (ODDS Ratio-1)

	RpR	RpY	YpR	YpY
upstream_10	0.1142	-0.1295	-0.1274	0.1150
upstream_9	0.1145	-0.1302	-0.1283	0.1167
upstream_8	0.1147	-0.1303	-0.1283	0.1165
upstream_7	0.1140	-0.1296	-0.1273	0.1162
upstream_6	0.1157	-0.1302	-0.1282	0.1151
upstream_5	0.1166	-0.1318	-0.1300	0.1165
upstream_4	0.1181	-0.1331	-0.1312	0.1176
upstream_3	0.1192	-0.1341	-0.1329	0.1195
upstream_2	0.1250	-0.1379	-0.1364	0.1250
upstream_1	0.1339	-0.1433	-0.1432	0.1351

REPEAT MASKED upstream datasets
Individual DINUCLEOTIDES: (ODDS Ratio-1)

	WpW	WpS	SpW	SpS
upstream_10	-0.0645	0.0782	0.0789	-0.1136
upstream_9	-0.0661	0.0805	0.0805	-0.1137
upstream_8	-0.0651	0.0797	0.0797	-0.1140
upstream_7	-0.0651	0.0792	0.0792	-0.1133
upstream_6	-0.0654	0.0784	0.0783	-0.1125
upstream_5	-0.0636	0.0784	0.0786	-0.1149
upstream_4	-0.0625	0.0767	0.0771	-0.1131
upstream_3	-0.0628	0.0760	0.0764	-0.1111
upstream_2	-0.0582	0.0695	0.0696	-0.0997
upstream_1	-0.0418	0.0373	0.0368	-0.0370

REPEAT MASKED upstream datasets

Individual DINUCLEOTIDES: (ODDS Ratio-1)

	MpM	MpK	KpM	KpK
upstream_10	0.0768	-0.0893	-0.0888	0.0783
upstream_9	0.0758	-0.0885	-0.0882	0.0776
upstream_8	0.0769	-0.0896	-0.0890	0.0790
upstream_7	0.0769	-0.0892	-0.0886	0.0785
upstream_6	0.0768	-0.0891	-0.0882	0.0780
upstream_5	0.0768	-0.0891	-0.0888	0.0786
upstream_4	0.0773	-0.0902	-0.0897	0.0792
upstream_3	0.0787	-0.0910	-0.0909	0.0803
upstream_2	0.0802	-0.0904	-0.0904	0.0808
upstream_1	0.0620	-0.0695	-0.0693	0.0619

Appendix C

C.1 Common patterns within ATCG upstream (unmasked) sequence

The following are tables and graphs of number common patterns in upstream sequences entire datasets each containing 18,725 sequences. The results show the changes in sequence similarity via a patterns analysis across the five upstream datasets; *upstream1*-to-*upstream5*. It is seen that sequence similarity decreases in the downstream directions and is markedly lower in *upstream1*. This is true when 20, 15 and 10 base common patterns are searched in the upstream datasets. The distinction in levels of pattern similarity becomes greater between the upstream datasets (*upstream1*-to-*upstream5*) the longer the pattern length. i.e. 20 base patterns show the best distinction in levels of similarity.

A/T/C/G: 20 BASE PATTERNS

Number of different patterns common to (x) number of upstream sequences

No of upstream Sequences (x) / upstream region	10	11	12	13	14	15	16	17	18	19
Upstream1	2245	1736	1394	1303	1432	773	716	612	532	543
Upstream2	3869	3142	2449	2154	2138	1645	1470	1328	1201	1416
Upstream3	4243	3866	2720	2392	2062	1851	1555	1502	1448	1666
Upstream4	4468	4077	3138	2598	2201	1848	1612	1435	1324	1925
Upstream5	4734	4537	3208	2569	2197	2055	1769	1835	1507	1275

No of upstream Sequences (x) / upstream region	20	21	22	23	24	25	26	27	28	29
Upstream1	535	580	680	411	294	277	222	242	183	175
Upstream2	875	800	703	678	592	543	539	458	406	379
Upstream3	919	908	820	773	646	599	576	444	498	396
Upstream4	1021	917	768	754	682	643	592	579	520	460
Upstream5	1068	1012	846	765	740	632	646	610	506	442

No of upstream Sequences (x) / upstream region	30	31	32	33	34	35	36	37	38	39
Upstream1	201	190	134	131	114	124	113	112	79	118
Upstream2	405	323	335	306	286	245	225	224	200	197
Upstream3	376	347	362	320	351	286	284	247	242	254
Upstream4	437	436	399	363	366	255	255	214	214	188
Upstream5	467	395	418	351	323	281	288	288	238	215

No of upstream Sequences (x) / upstream region	40	41	42	43	44	45	46	47	48	49
Upstream1	110	109	104	90	74	108	76	55	53	47
Upstream2	189	174	171	162	144	129	128	110	121	103
Upstream3	199	220	178	169	175	183	149	127	140	129
Upstream4	207	215	178	162	137	141	138	140	142	135
Upstream5	198	205	212	204	158	161	166	146	134	119

No of upstream Sequences (x) / upstream region	50	51	52	53	54	55	56	57	58	59
Upstream1	38	46	38	23	50	36	45	41	26	41
Upstream2	105	92	88	99	104	93	84	77	63	62
Upstream3	111	95	103	94	128	114	90	78	85	79
Upstream4	114	122	117	113	71	88	99	83	101	98
Upstream5	93	122	107	120	106	113	100	88	90	77

A/T/C/G: 15 BASE PATTERNS

Number of different patterns common to (x) number of upstream sequences

No of upstream Sequences (x) / upstream region	10	11	12	13	14	15	16	17	18	19
Upstream1	3227	2623	1861	1731	1858	1137	1040	960	761	783
Upstream2	4829	3951	2933	2684	2388	1890	1639	1582	1440	1680
Upstream3	5148	4699	3316	2776	2503	2053	1832	1803	1674	2025
Upstream4	5563	4761	3500	3023	2659	2247	1994	1701	1543	2136
Upstream5	5666	5093	3810	3155	2722	2361	2107	2025	1784	1562

No of upstream Sequences (x) / upstream region	20	21	22	23	24	25	26	27	28	29
Upstream1	724	737	813	582	483	407	367	306	300	304
Upstream2	1163	939	833	802	707	647	610	616	636	520
Upstream3	1237	1107	985	839	802	674	666	597	656	594
Upstream4	1218	1106	972	936	828	815	774	664	631	592
Upstream5	1287	1179	1080	1006	901	798	798	709	683	625

No of upstream Sequences (x) / upstream region	30	31	32	33	34	35	36	37	38	39
Upstream1	275	206	239	223	217	170	208	189	147	168
Upstream2	475	414	413	399	358	337	336	342	256	298
Upstream3	521	513	525	438	419	395	318	312	306	297
Upstream4	535	468	489	419	435	356	349	353	353	326
Upstream5	632	541	533	517	446	411	361	332	330	312

No of upstream Sequences (x) / upstream region	40	41	42	43	44	45	46	47	48	49
Upstream1	151	139	150	120	95	133	112	102	82	79
Upstream2	261	250	205	200	210	186	154	176	160	149
Upstream3	278	261	249	273	206	207	167	167	159	154
Upstream4	303	291	229	222	225	194	191	200	166	172
Upstream5	289	257	247	258	259	202	218	208	192	155

No of upstream Sequences (x) / upstream region	50	51	52	53	54	55	56	57	58	59
Upstream1	70	87	62	56	58	65	68	60	58	45
Upstream2	136	146	136	124	118	114	108	109	122	113
Upstream3	181	182	161	131	134	117	108	113	109	98
Upstream4	156	141	158	135	147	152	119	111	99	111
Upstream5	153	120	143	124	132	139	133	124	125	127

A/T/C/G: 10 BASE PATTERNS

Number of different patterns common to (x) number of upstream sequences

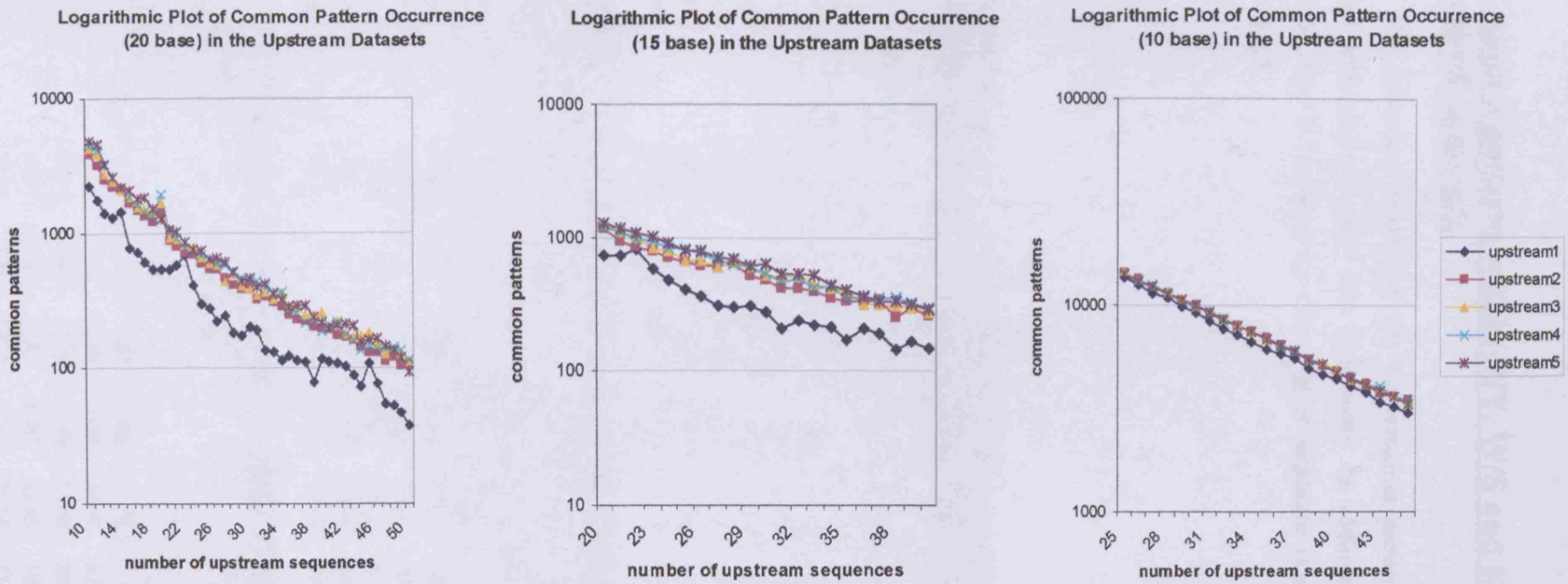
No of upstream Sequences (x) / upstream region	10	11	12	13	14	15	16	17	18	19
Upstream1	38734	36696	34941	33081	30940	29221	27271	25342	23364	21851
Upstream2	27277	26595	25903	25704	25037	24436	23743	22666	21629	20585
Upstream3	24613	24347	24086	24000	23950	23493	22857	21875	20876	20482
Upstream4	23966	24046	23939	23730	23407	23154	22577	21984	21062	19910
Upstream5	23908	23990	24261	23664	23377	23349	22415	21718	21039	19941

No of upstream Sequences (x) / upstream region	20	21	22	23	24	25	26	27	28	29
Upstream1	20042	18417	17051	15526	14555	13620	12602	11344	10786	9785
Upstream2	19482	18381	17299	16242	15363	14310	13338	12251	11396	10576
Upstream3	19037	18211	16917	15993	15056	14499	13199	12298	11538	10699
Upstream4	19108	18167	17159	16061	15032	14185	13112	12511	11439	10659
Upstream5	19025	18148	17230	15734	14841	14044	13046	12430	11212	10714

No of upstream Sequences (x) / upstream region	30	31	32	33	34	35	36	37	38	39
Upstream1	9146	8310	7741	7200	6695	6116	5862	5548	4988	4654
Upstream2	9962	9137	8519	7965	7501	7010	6403	5898	5485	5096
Upstream3	9861	9367	8586	8060	7509	6839	6440	5986	5489	5260
Upstream4	9736	9351	8570	7994	7483	6917	6287	5897	5526	5215
Upstream5	9963	9185	8562	8002	7489	6915	6404	5993	5447	5207

No of upstream Sequences (x) / upstream region	40	41	42	43	44	45	46	47	48	49
Upstream1	4393	4063	3785	3425	3240	3031	2906	2743	2474	2354
Upstream2	4766	4505	4199	3779	3652	3489	3158	2978	2707	2620
Upstream3	4903	4368	4150	3962	3694	3402	3194	2999	2856	2463
Upstream4	4754	4457	4167	4082	3659	3537	3160	2974	2715	2630
Upstream5	4750	4431	4169	3868	3662	3359	3176	2918	2847	2586

No of upstream Sequences (x) / upstream region	50	51	52	53	54	55	56	57	58	59
Upstream1	2283	2058	2023	1863	1779	1695	1521	1481	1384	1374
Upstream2	2444	2325	2122	1907	1885	1790	1608	1525	1524	1248
Upstream3	2452	2274	2076	2037	1934	1822	1656	1563	1536	1464
Upstream4	2508	2279	2130	2029	1886	1819	1646	1539	1514	1341
Upstream5	2447	2229	2193	2010	1836	1741	1641	1589	1513	1339



Graphs showing an analysis of common patterns within each of the five 1Kb upstream datasets: *upstream1-to-upstream5*. The results show that common patterns become lower in the downstream direction and that *upstream1* contains the lowest level of common patterns. This means that in the upstream sequence towards the TSS, sequence similarity decreases . This is true when 20, 15, and 10 base common patterns are determined.

C.2 Common patterns within R/Y, W/S and K/M upstream (unmasked) sequence

The following data-tables are for common patterns (20 base) found in upstream sequences translated into R/Y and W/S bases. In addition results are shown for an K/M translation. Here an A and C in the original sequence is converted to M, and T and G is converted to K.

R/Y: 20 BASE PATTERNS

Number of different patterns common to (x) number of upstream sequences

No of upstream Sequences (x) / upstream region	20	21	22	23	24	25	26	27	28	29
Upstream1	21278	19011	17076	15124	14016	12369	11136	10167	9307	8362
Upstream2	19212	17165	15209	13618	12344	10913	9860	8977	8203	7389
Upstream3	19147	16737	14929	13330	11974	10751	9415	8575	7863	7125
Upstream4	18854	16654	14906	12985	11857	10803	9661	8579	7651	7057
Upstream5	18685	16550	14678	13042	11758	10336	9552	8556	7662	6984

No of upstream Sequences (x) / upstream region	30	31	32	33	34	35	36	37	38	39
Upstream1	7521	7027	6370	5611	5357	4870	4574	4127	3717	3437
Upstream2	6646	6073	5570	4962	4692	4255	3973	3662	3432	3069
Upstream3	6528	5938	5285	4840	4460	4122	3895	3513	3204	3118
Upstream4	6266	5819	5382	4796	4330	4031	3755	3423	3114	2926
Upstream5	6223	5832	5218	4878	4410	3951	3630	3386	3171	2898

No of upstream Sequences (x) / upstream region	40	41	42	43	44	45	46	47	48	49
Upstream1	3311	3031	2845	2684	2460	2302	2194	2064	1956	1830
Upstream2	2881	2698	2470	2309	2198	2043	1945	1793	1694	1559
Upstream3	2799	2493	2333	2293	2135	1904	1853	1686	1609	1510
Upstream4	2809	2556	2443	2242	2114	2125	1882	1657	1598	1532
Upstream5	2730	2511	2314	2203	2109	1991	1807	1770	1638	1564

No of upstream Sequences (x) / upstream region	50	51	52	53	54	55	56	57	58	59
Upstream1	1690	1560	1513	1422	1345	1188	1142	1137	1045	1051
Upstream2	1533	1352	1368	1308	1221	1164	1082	1028	922	898
Upstream3	1406	1405	1310	1236	1150	1087	999	937	921	828
Upstream4	1451	1323	1236	1240	1178	1135	1098	1005	951	867
Upstream5	1407	1428	1322	1238	1127	1083	1049	967	916	872

No of upstream Sequences (x) / upstream region	60	61	62	63	64	65	66	67	68	69
Upstream1	949	877	883	792	735	703	669	640	651	592
Upstream2	851	804	758	687	640	644	656	593	572	489
Upstream3	854	790	725	672	671	663	621	610	525	470
Upstream4	827	757	735	718	648	626	631	585	529	528
Upstream5	816	809	764	689	655	633	649	538	568	556

W/S: 20 BASE PATTERNS

Number of different patterns common to (x) number of upstream sequences

No of upstream Sequences (x) / upstream region	20	21	22	23	24	25	26	27	28	29
Upstream1	20358	18054	15876	13984	12647	11131	10079	9047	7998	7320
Upstream2	20977	19176	17100	15296	13866	12599	11306	10334	9189	8422
Upstream3	19928	18152	16230	14780	13532	12181	11113	9924	9323	8316
Upstream4	19998	17718	15967	14564	13337	11963	10773	10080	8872	8150
Upstream5	19735	17805	16186	14245	13116	11833	10849	9701	9051	8125

No of upstream Sequences (x) / upstream region	30	31	32	33	34	35	36	37	38	39
Upstream1	6657	5907	5520	4913	4591	4155	3745	3559	3284	3118
Upstream2	7591	6829	6278	5789	5374	4838	4503	4150	3752	3606
Upstream3	7604	7061	6403	5836	5428	5103	4555	4329	3958	3713
Upstream4	7621	6936	6271	5989	5389	4873	4634	4290	3937	3726
Upstream5	7568	6798	6318	6004	5283	4912	4635	4280	3808	3675

No of upstream Sequences (x) / upstream region	40	41	42	43	44	45	46	47	48	49
Upstream1	2826	2793	2530	2400	2214	2165	1975	1851	1791	1750
Upstream2	3293	3018	2854	2662	2498	2361	2183	2044	1906	1682
Upstream3	3390	3080	2868	2699	2427	2353	2203	2052	1908	1719
Upstream4	3369	3081	2861	2635	2457	2300	2226	2078	1998	1827
Upstream5	3448	3148	2968	2820	2577	2379	2219	2143	2008	1811

No of upstream Sequences (x) / upstream region	50	51	52	53	54	55	56	57	58	59
Upstream1	1611	1483	1357	1307	1191	1182	1074	1052	957	949
Upstream2	1737	1646	1549	1439	1375	1272	1234	1174	1092	1099
Upstream3	1644	1633	1570	1466	1381	1362	1272	1202	1120	1154
Upstream4	1774	1681	1523	1463	1387	1333	1235	1191	1119	1069
Upstream5	1744	1647	1563	1447	1329	1336	1278	1203	1145	1143

No of upstream Sequences (x) / upstream region	60	61	62	63	64	65	66	67	68	69
Upstream1	809	803	697	692	697	651	623	606	555	540
Upstream2	1035	988	993	969	900	831	878	762	751	675
Upstream3	1092	990	989	954	867	830	814	762	785	753
Upstream4	1047	1042	984	884	885	834	784	756	729	707
Upstream5	1056	1034	1041	959	927	882	858	828	786	702

K/M: 20 BASE PATTERNS

Number of different patterns common to (x) number of upstream sequences

No of upstream Sequences (x) / upstream region	20	21	22	23	24	25	26	27	28	29
Upstream1	39773	34664	30152	26094	22289	18950	16477	13856	11973	9965
Upstream2	33449	29246	26028	22503	19860	17083	14839	13051	11248	9706
Upstream3	31866	28174	24534	21800	18973	16289	14300	12486	10858	9371
Upstream4	31760	28261	24533	21554	18765	16573	14318	12332	10763	9387
Upstream5	31684	27832	24849	21533	18533	16218	14413	12332	10929	9355

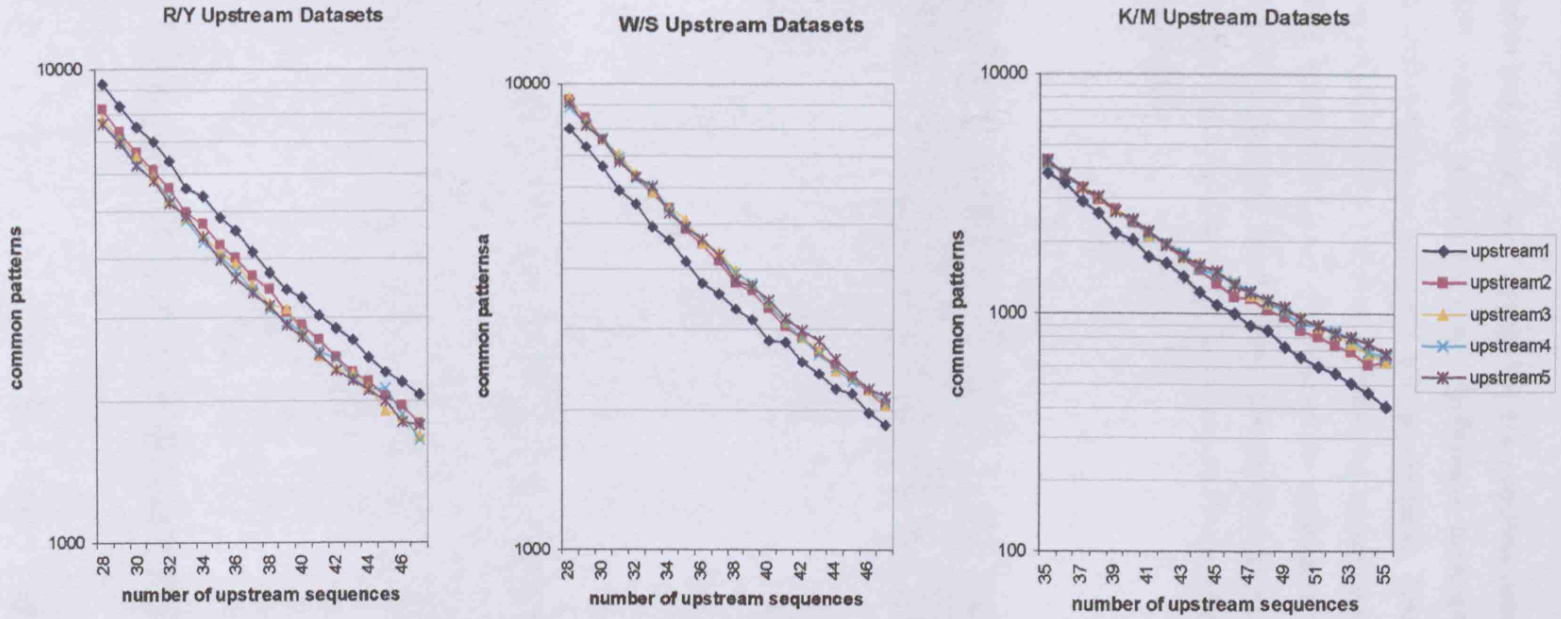
No of upstream Sequences (x) / upstream region	30	31	32	33	34	35	36	37	38	39
Upstream1	8529	7409	6166	5401	4562	3834	3479	2926	2599	2166
Upstream2	8538	7549	6396	5684	5032	4322	3788	3323	2967	2685
Upstream3	8252	7142	6323	5572	4939	4254	3829	3367	3045	2711
Upstream4	8275	7230	6404	5589	4826	4203	3805	3324	3058	2668
Upstream5	8104	7143	6352	5526	4935	4327	3767	3306	3086	2636

No of upstream Sequences (x) / upstream region	40	41	42	43	44	45	46	47	48	49
Upstream1	2017	1722	1616	1422	1235	1091	999	895	852	724
Upstream2	2369	2139	1934	1703	1514	1344	1180	1154	1040	972
Upstream3	2444	2106	1947	1712	1624	1502	1343	1198	1095	1042
Upstream4	2397	2139	1954	1774	1532	1451	1351	1245	1125	1020
Upstream5	2443	2174	1938	1717	1579	1485	1304	1227	1133	1065

No of upstream Sequences (x) / upstream region	50	51	52	53	54	55	56	57	58	59
Upstream1	665	601	558	512	461	406	390	376	375	323
Upstream2	854	802	737	687	613	625	555	517	486	469
Upstream3	974	880	834	747	686	628	604	541	557	506
Upstream4	904	888	856	782	694	664	623	548	522	491
Upstream5	947	890	828	795	752	686	622	595	546	519

No of upstream Sequences (x) / upstream region	60	61	62	63	64	65	66	67	68	69
Upstream1	319	280	289	268	242	196	226	209	186	179
Upstream2	426	394	399	433	382	343	334	294	288	252
Upstream3	448	442	448	388	373	370	307	333	259	274
Upstream4	509	454	423	392	353	342	366	301	295	322
Upstream5	472	478	399	396	398	359	358	296	279	283

Sequence similarity in Unmasked Upstream Sequence



Three graphs showing the results of a common patterns analysis for the five 1Kb upstream datasets; *upstream1-to-upstream5*. Graph A shows common patterns for R/Y (translated) sequences, graph B for W/S sequences and graph C for K/M sequences. K/M sequences refer to a conversion of A and C into 'M' and T and G into 'K'.
The results show that when the sequence is viewed from the W/S and K/M perspective common patterns decrease from *upstream5* to *upstream1* (in the downstream direction). From the R/Y perspective the opposite is true. This means that for R/Y the sequence becomes more similar across the 5Kb region towards the TSS. In contrast for W/S and K/M the sequence becomes less similar. The result suggests that purines and pyrimidines associate to form sequence in a more specific way towards the TSS than the equivalent W/S base association. The K/M sequence is utilised here as a control, showing an association that is neither R/Y or W/S.

C.3 Common patterns within ATCG upstream sequence: REPEAT MASKED

The table and graph show results for the common pattern analysis on repeat masked sequences. The results reveal no clear distinction between the five upstream datasets (*upstream1*-to-*upstream5*) with respect to pattern similarity. This is due to noise. Therefore it is not possible to conclude that there is a difference in sequence similarity for the upstream positional repeat masked datasets. This leaves an uncertainty as to whether there is a true difference between equivalent masked and unmasked sequence datasets. The R/Y, W/S and K/M datasets thought proved less noisy and therefore it was possible to make such comparisons (see sections that follow).

A/T/C/G: 20 BASE PATTERNS
REPEAT MASKED SEQUENCES
Number of different patterns common to (x) number of upstream sequences

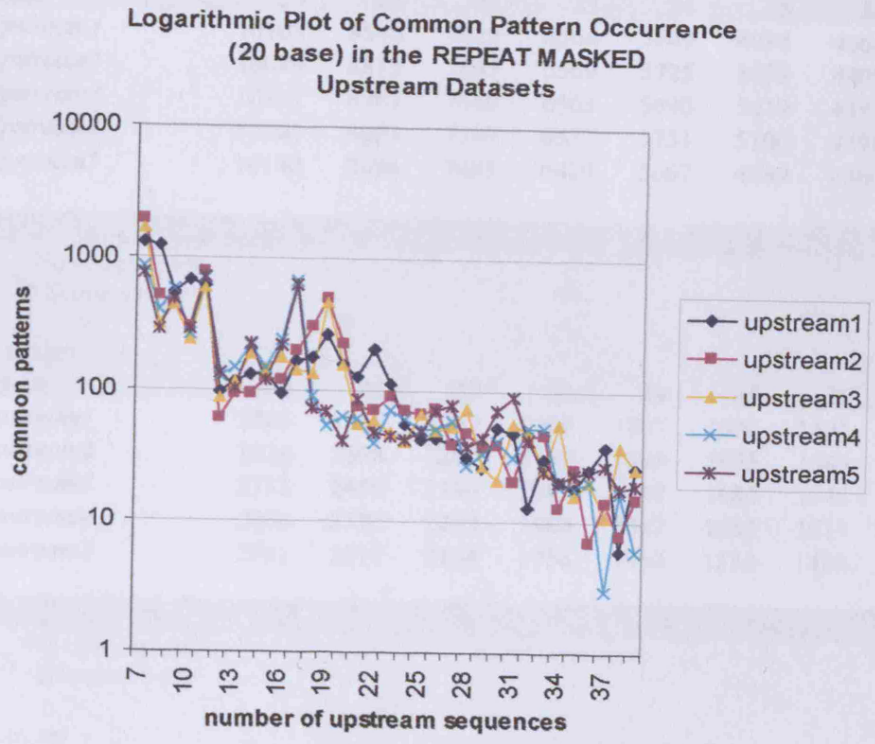
No of upstream Sequences (x) / upstream region	10	11	12	13	14	15	16	17	18	19
Upstream1	675	615	96	115	129	134	93	165	178	264
Upstream2	292	775	61	94	94	121	120	201	302	487
Upstream3	241	592	86	115	188	144	173	144	134	463
Upstream4	258	697	126	146	219	145	236	660	91	54
Upstream5	286	714	134	102	226	120	216	622	73	68

No of upstream Sequences (x) / upstream region	20	21	22	23	24	25	26	27	28	29
Upstream1	156	125	205	111	56	42	44	39	31	26
Upstream2	227	74	70	91	71	66	72	38	48	37
Upstream3	157	56	59	48	41	66	49	54	73	27
Upstream4	62	60	39	71	60	50	50	58	27	33
Upstream5	41	84	47	45	42	49	70	78	37	44

No of upstream Sequences (x) / upstream region	30	31	32	33	34	35	36	37	38	39
Upstream1	52	47	13	30	20	18	25	37	6	26
Upstream2	38	20	45	46	13	25	7	14	8	15
Upstream3	21	56	49	25	56	16	20	11	37	26
Upstream4	41	31	52	52	19	18	20	3	17	6
Upstream5	71	90	42	24	20	23	24	27	18	19

No of upstream Sequences (x) / upstream region	40	41	42	43	44	45	46	47	48	49
Upstream1	31	31	35	12	12	44	17	1	0	0
Upstream2	9	10	19	11	4	13	5	1	1	1
Upstream3	13	18	21	7	21	14	1	0	0	0
Upstream4	12	7	5	5	10	1	0	1	2	0
Upstream5	8	9	15	5	0	0	0	0	0	0

No of upstream Sequences (x) / upstream region	50	51	52	53	54	55	56	57	58	59
Upstream1	0	0	0	0	0	0	0	0	0	0
Upstream2	0	0	0	0	0	0	0	0	0	0
Upstream3	0	0	0	0	0	0	0	0	0	0
Upstream4	0	0	0	0	0	0	0	0	0	0
Upstream5	0	0	0	0	0	0	0	0	0	0



Graphs showing an analysis of common patterns within each of the five 1Kb repeat masked upstream datasets: *upstream1-to-upstream5*. The results are noisy and show no obvious distinction in common patterns levels between these five datasets.

C.4 Common patterns within R/Y, W/S and K/M upstream sequence: REPEAT MASKED

The following tables and graphs show common pattern results for the five upstream datasets (upstream1-to-upstream5). The datasets include sequences translated into R/Y, W/S and K/M each for which common patterns were analysed. The results revealed that the repeat masked sequences display similar trends in relative sequence similarity across the 5Kb upstream as the unmasked sequences. In the downstream direction (and particularly for uptream1) there is increased R/Y sequence similarity. In contrast, for W/S and K/M there is decreased sequence similarity in the downstream direction. Therefore the overall trends across the upstream are the same for repeat masked and unmasked sequences.

R/Y: 20 BASE PATTERNS
REPEAT MASKED SEQUENCES
Number of different patterns common to (x) number of upstream sequences

No of upstream Sequences (x) / upstream region	20	21	22	23	24	25	26	27	28	29
Upstream1	10103	8540	7429	6564	5749	4998	4660	4060	3651	3230
Upstream2	10079	8815	7647	6569	5725	5079	4405	4009	3538	3100
Upstream3	10207	8783	7640	6563	5690	5019	4387	3800	3422	2990
Upstream4	10161	8893	7587	6533	5751	5106	4498	3873	3535	3189
Upstream5	10140	8484	7481	6420	5667	4889	4360	3988	3402	3081

No of upstream Sequences (x) / upstream region	30	31	32	33	34	35	36	37	38	39
Upstream1	2863	2615	2273	2069	1887	1800	1591	1456	1378	1226
Upstream2	2824	2598	2269	2130	1889	1805	1585	1421	1283	1141
Upstream3	2731	2459	2194	2001	1740	1680	1548	1394	1164	1126
Upstream4	2806	2372	2202	1989	1837	1657	1513	1346	1225	1091
Upstream5	2791	2512	2164	1956	1830	1556	1489	1357	1277	1115

No of upstream Sequences (x) / upstream region	40	41	42	43	44	45	46	47	48	49
Upstream1	1117	1040	977	839	768	725	664	643	522	547
Upstream2	1088	1011	849	786	720	664	623	572	517	519
Upstream3	1061	933	899	812	703	623	636	543	523	497
Upstream4	1035	939	791	740	658	611	536	546	504	445
Upstream5	940	909	827	792	684	686	559	561	471	479

No of upstream Sequences (x) / upstream region	50	51	52	53	54	55	56	57	58	59
Upstream1	510	465	414	372	359	324	304	287	288	280
Upstream2	437	418	394	364	319	346	261	307	276	240
Upstream3	456	378	419	380	342	315	292	298	244	263
Upstream4	402	393	389	326	316	287	289	289	267	248
Upstream5	402	341	344	330	327	303	273	282	262	233

No of upstream Sequences (x) / upstream region	60	61	62	63	64	65	66	67	68	69
Upstream1	212	219	228	221	223	210	164	187	160	167
Upstream2	272	225	231	215	182	194	174	175	168	144
Upstream3	238	236	174	174	196	201	186	166	157	158
Upstream4	199	205	228	196	200	172	174	139	159	164
Upstream5	251	196	211	193	202	173	151	140	159	148

W/S: 20 BASE PATTERNS
REPEAT MASKED SEQUENCES

Number of different patterns common to (x) number of upstream sequences

No of upstream Sequences (x) / upstream region	20	21	22	23	24	25	26	27	28	29
Upstream1	8297	7064	6261	5477	4841	4288	3949	3394	2991	2798
Upstream2	10346	8900	7661	6459	5804	5010	4308	3827	3356	3058
Upstream3	10109	8880	7814	6712	5860	5098	4452	4037	3637	3083
Upstream4	10244	8994	7614	6696	5864	5091	4381	4000	3592	3193
Upstream5	10291	8892	7638	6723	5723	5128	4515	3908	3557	3147

No of upstream Sequences (x) / upstream region	30	31	32	33	34	35	36	37	38	39
Upstream1	2483	2179	1977	1849	1688	1492	1378	1339	1187	1110
Upstream2	2711	2463	2282	2164	1976	1850	1774	1686	1525	1377
Upstream3	2860	2503	2357	2169	1995	1891	1720	1650	1454	1407
Upstream4	2871	2625	2431	2227	1994	1929	1777	1639	1537	1447
Upstream5	2789	2586	2417	2145	1992	1833	1732	1633	1499	1438

No of upstream Sequences (x) / upstream region	40	41	42	43	44	45	46	47	48	49
Upstream1	956	912	860	789	801	737	736	727	646	639
Upstream2	1205	1258	1104	977	899	852	775	696	623	590
Upstream3	1349	1213	1190	1092	989	935	821	747	652	651
Upstream4	1352	1222	1226	1167	975	953	893	809	706	664
Upstream5	1373	1261	1169	1084	1027	972	874	798	770	649

No of upstream Sequences (x) / upstream region	50	51	52	53	54	55	56	57	58	59
Upstream1	580	609	590	487	472	494	474	438	438	367
Upstream2	508	477	418	380	365	299	319	275	300	248
Upstream3	575	545	456	429	418	343	337	268	282	273
Upstream4	600	552	511	435	406	392	317	302	277	235
Upstream5	625	563	549	447	415	375	341	338	299	285

No of upstream Sequences (x) / upstream region	60	61	62	63	64	65	66	67	68	69
Upstream1	409	337	339	292	274	278	226	227	207	195
Upstream2	229	235	199	248	210	195	204	225	210	170
Upstream3	232	186	207	195	182	162	176	161	151	183
Upstream4	216	233	190	183	178	165	157	140	146	174
Upstream5	270	230	206	207	184	220	171	165	182	168

K/M: 20 BASE PATTERNS REPEAT MASKED SEQUENCES

Number of different patterns common to (x) number of upstream sequences

No of upstream Sequences (x) / upstream region	20	21	22	23	24	25	26	27	28	29
Upstream1	9188	7050	5256	4120	3155	2500	1973	1539	1290	1028
Upstream2	10572	8270	6599	5085	3987	3242	2558	1990	1557	1347
Upstream3	10798	8544	6570	5184	4046	3186	2585	1931	1654	1259
Upstream4	10872	8429	6641	5080	4019	3204	2480	2018	1517	1305
Upstream5	10731	8251	6612	5097	4006	3235	2480	1974	1607	1338

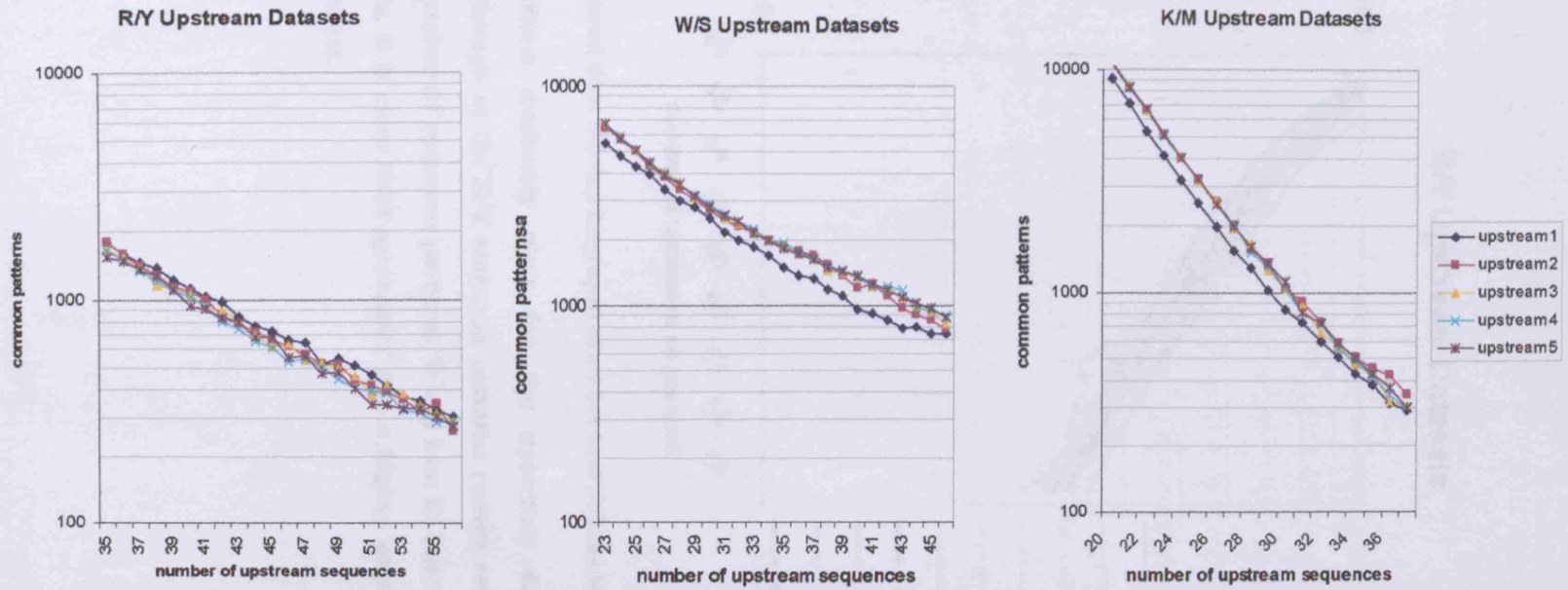
No of upstream Sequences (x) / upstream region	30	31	32	33	34	35	36	37	38	39
Upstream1	830	731	594	508	426	379	315	294	247	238
Upstream2	1035	906	731	591	510	455	422	344	286	271
Upstream3	1044	875	648	573	474	435	321	306	239	221
Upstream4	1039	813	717	556	488	418	346	297	242	217
Upstream5	1123	809	719	586	500	429	371	303	295	239

No of upstream Sequences (x) / upstream region	40	41	42	43	44	45	46	47	48	49
Upstream1	220	183	181	114	120	103	85	112	89	93
Upstream2	211	175	151	147	118	113	107	86	101	75
Upstream3	164	166	175	145	110	109	113	108	122	96
Upstream4	197	189	131	130	115	94	99	85	87	71
Upstream5	209	174	170	153	140	124	115	90	95	91

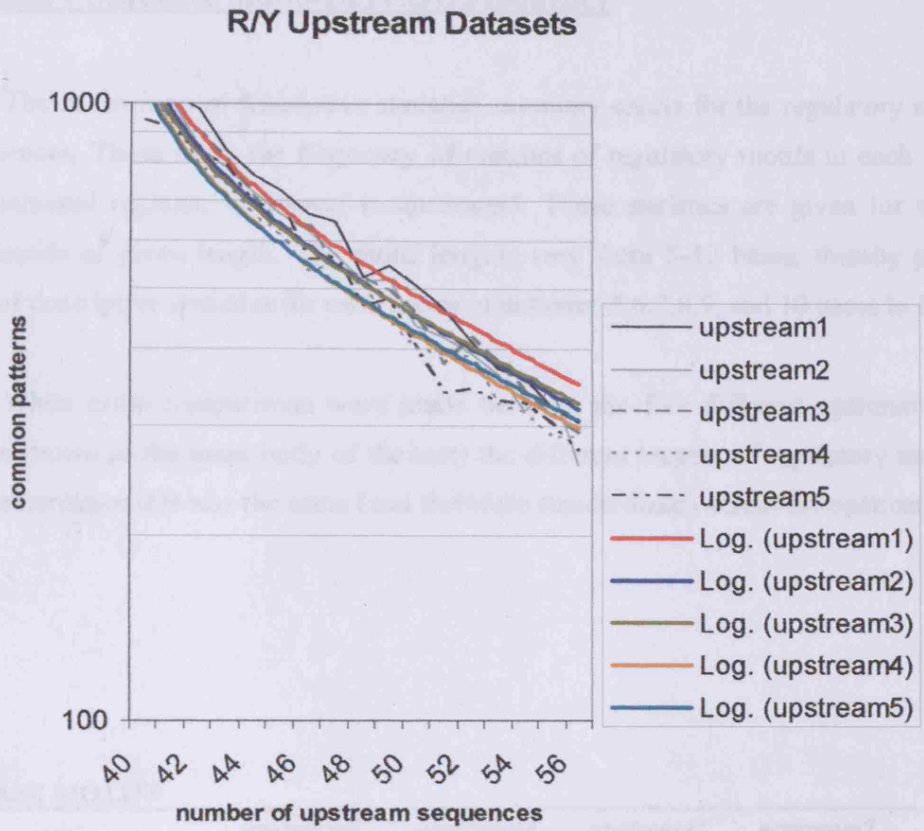
No of upstream Sequences (x) / upstream region	50	51	52	53	54	55	56	57	58	59
Upstream1	78	68	60	66	54	65	43	50	39	38
Upstream2	75	68	70	57	52	54	54	34	49	43
Upstream3	79	85	58	61	53	60	48	41	36	46
Upstream4	69	73	47	59	62	35	48	38	43	28
Upstream5	82	73	68	44	62	47	30	38	43	24

No of upstream Sequences (x) / upstream region	60	61	62	63	64	65	66	67	68	69
Upstream1	26	21	32	22	21	30	22	17	20	13
Upstream2	33	32	37	27	30	28	22	18	12	11
Upstream3	41	34	20	22	18	18	18	16	17	13
Upstream4	42	40	31	19	26	36	23	28	19	15
Upstream5	32	28	26	21	21	19	23	18	7	13

Sequence similarity in Masked Upstream Sequence



Three graphs showing the results of a common patterns analysis for the five 1Kb repeat masked upstream datasets; *upstream1-to-upstream5*.
Graph A shows common patterns for R/Y (translated) sequences, graph B for W/S sequences and graph C for K/M sequences. K/M sequences refer to a conversion of A and C into 'M' and T and G into 'K'.
Results reveal that when the sequence is viewed from the W/S and K/M perspective common patterns decrease from *upstream5* to *upstream1* (in the downstream direction). From the R/Y perspective the opposite is true. This means that for R/Y the sequence becomes more similar across the 5Kb region towards the TSS. However, the R/Y results are noisy.
These results indicate that the trends for sequence similarity (for R/Y, W/S and K/M) are the same for repeat masked upstream sequences as they are for the equivalent unmasked sequences.



***Note; the coloured lines labeled Log(upstream) are best fit lines for the logarithmic plots.**

The R/Y pattern similarity plots for the upstream datasets are noisy. A closer examination though of the R/Y common patterns results reveals that *upstream1* contains the highest number of common patterns When best fit lines are plotted for the upstream dataset results, it is clear that *upstream1* has a higher pattern similarity than the other upstream datasets.

Appendix D

D.1 Regulatory Sequence Motifs (ATCG): Frequency

The following are descriptive statistics summary charts for the regulatory motif set of sample sequences. These show the frequency of matches of regulatory motifs in each of the 1Kb upstream positional regions; *upstream1*-to-*upstream5*. These statistics are given for each set of regulatory motifs of given length. The motif lengths vary from 5-10 bases, thereby providing a breakdown of descriptive statistics for each group of patterns; 5,6,7,8,9, and 10 bases in length.

When cross-comparisons were made between the five different upstream positional segments (as shown in the main body of the text) the different lengths of regulatory motifs where grouped together since this was the same (and therefore standardized) across the upstream datasets.

FIVE BASE MOTIFS

	upstream5	*upstream4*	*upstream3*	*upstream2*	*upstream1*
Mean	17261	17434	17448	17880	21528
Standard Error	4666	4738	4787	4890	5696
Median	13656	13784	13537	13174	13461
Standard Deviation	13997	14214	14360	14671	17087
Skewness	0.819	0.822	0.880	0.963	1.430
Count	9	9	9	9	9
Confidence Level(95.0%)	10759	10926	11038	11277	13134

SIX BASE MOTIFS

	upstream5	*upstream4*	*upstream3*	*upstream2*	*upstream1*
Mean	4397.59	4424.17	4460.10	4743.10	7096.17
Standard Error	720.59	729.44	729.13	734.16	1095.01
Median	3320	3269	3438	3719	4413
Standard Deviation	3880.49	3928.13	3926.50	3953.58	5896.79
Skewness	0.804	0.830	0.823	0.740	1.154
Count	29	29	29	29	29
Confidence Level(95.0%)	1476.06	1494.18	1493.56	1503.86	2243.02

SEVEN BASE MOTIFS

	upstream5	*upstream4*	*upstream3*	*upstream2*	*upstream1*
Mean	981.69	1014.44	1010.56	1029.63	1516.75

Standard Error	191.37	199.56	203.51	201.94	444.06
Median	916	888	908	903	1153
Standard Deviation	765.48	798.23	814.03	807.77	1776.26
Skewness	1.416	1.532	1.762	2.073	3.351
Count	16	16	16	16	16
Confidence Level(95.0%)	407.89	425.34	433.77	430.43	946.50

EIGHT BASE MOTIFS

	upstream5	*upstream4*	*upstream3*	*upstream2*	*upstream1*
Mean	306.15	308.83	315.34	340.75	668.05
Standard Error	54.05	54.76	54.74	56.45	131.41
Median	102	92	99	136	354
Standard Deviation	415.14	420.62	420.45	433.59	1009.42
Skewness	2.609	2.648	2.414	2.235	3.678
Count	59	59	59	59	59
Confidence Level(95.0%)	108.19	109.61	109.57	112.99	263.06

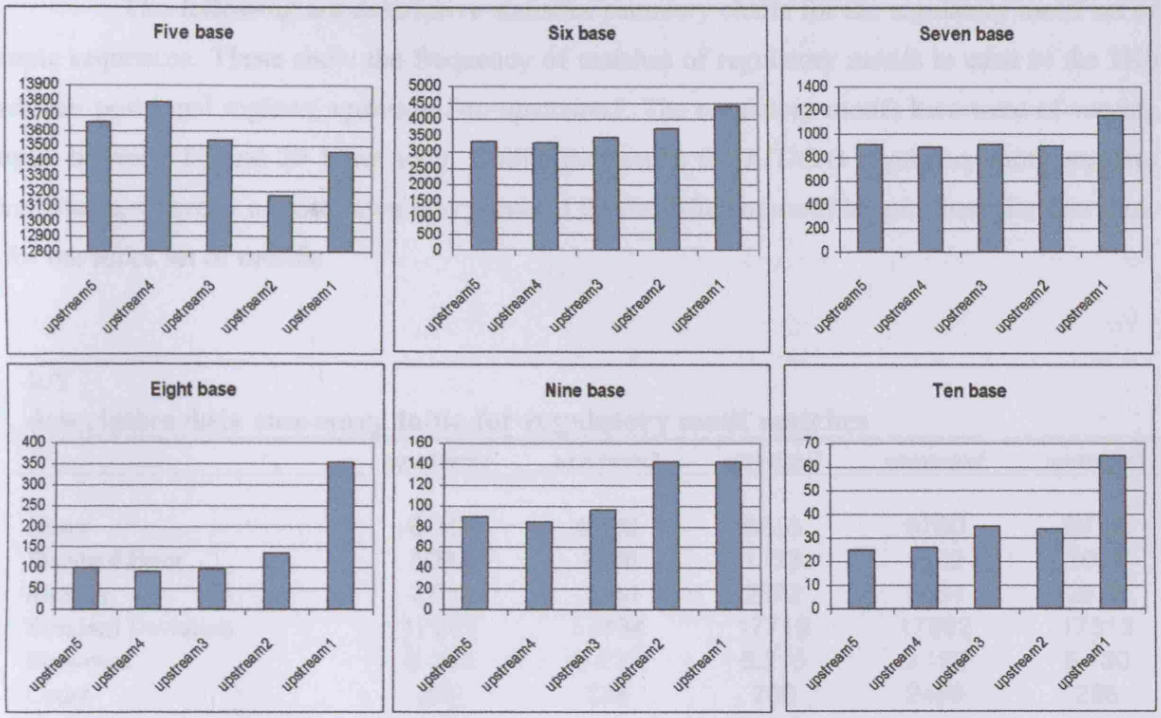
NINE BASE MOTIFS

	upstream5	*upstream4*	*upstream3*	*upstream2*	*upstream1*
Mean	222.29	233.88	240.29	286.88	660.65
Standard Error	73.38	77.75	79.41	89.34	239.38
Median	90	84	96	142	142
Standard Deviation	302.55	320.57	327.41	368.38	987.00
Skewness	2.201	2.082	2.180	1.666	2.643
Count	17	17	17	17	17
Confidence Level(95.0%)	155.55	164.82	168.34	189.40	507.47

TEN BASE MOTIFS

	upstream5	*upstream4*	*upstream3*	*upstream2*	*upstream1*
Mean	73.17	71.57	78.17	94.65	348.48
Standard Error	19.76	19.97	20.08	24.63	128.16
Median	25	26	35	34	65
Standard Deviation	94.75	95.76	96.31	118.10	614.64
Skewness	1.423	1.633	1.469	1.293	2.380
Count	23	23	23	23	23
Confidence Level(95.0%)	40.97	41.41	41.65	51.07	265.79

Frequency of Regulatory Motif Matches Across the Upstream Positional Regions



Frequency of Regulatory Motif Matches Across the Upstream Positional regions (*upstream5*-to-*upstream1*)						
Frequency: median values for each set of regulatory motifs						
Motif length / **Upstream region**	**ten**	**nine**	**eight**	**seven**	**six**	**five**
upstream5	25	90	102	915.5	3320	13656
upstream4	26	84	92	888	3269	13784
upstream3	35	96	99	907.5	3438	13537
upstream2	34	142	136	903	3719	13174
upstream1	65	142	354	1153	4413	13461

The plots and data-table above show the frequency of regulatory motif matches across the upstream datasets. Each plot contains the number of matches of regulatory motifs of specified length (in bases) within each upstream dataset. The median number of matches is given across the dataset of motifs. For all the motif lengths, except for five base, the frequency of matches increases on average from *upstream5*-to-*upstream1* in the downstream direction. *Upstream1* is the most distinctive for motif matches.

D.2 Regulatory Sequence Motifs (R/Y and W/S): Frequency

The following are descriptive statistics summary charts for the regulatory motif set of sample sequences. These show the frequency of matches of regulatory motifs in each of the 1Kb upstream positional regions; *upstream1*-to-*upstream5*. The regulatory motifs here were of varying length, between 10 and 20 bases long. Unlike the results for A/T/C/G regulatory motif matches given above, whereby a breakdown was provided for the different motif lengths, here the data show is for the entire set of motifs.

R/Y

descriptive data summary table for regulatory motif matches

	upstream1	*upstream2*	*upstream3*	*upstream4*	*upstream5*
Mean	6829	6799	6816	6760	6757
Standard Error	1094	1108	1123	1102	1097
Median	2702	2600	2672	2664	2676
Standard Deviation	17265	17484	17719	17392	17313
Skewness	6.042	6.137	6.215	6.192	6.130
Count	286	286	286	2486	286
Confidence Level(95.0%)	2155	2182	2212	2171	2161

W/S

descriptive data summary table for regulatory motif matches

	upstream1	*upstream2*	*upstream3*	*upstream4*	*upstream5*
Mean	7986	3444	2992	2889	2894
Standard Error	1346	362	310	299	300
Median	2688	1548	1421	1365	1333
Standard Deviation	20157	5425	4644	4467	4496
Skewness	6.793	3.138	3.034	3.042	3.042
Count	224	224	224	224	224
Confidence Level(95.0%)	2654	714	612	589	592

D.3 Regulatory Sequence Motifs (A/T/C/G): Representation

The following are descriptive statistics summary charts for the regulatory motif set of sample sequences. These show the representation of regulatory motifs in each of the 1Kb upstream positional regions; *upstream1*-to-*upstream5*. These statistics are given for each set of regulatory motifs of given length. The motif lengths vary from 5-10 bases, thereby providing a breakdown of descriptive statistics for each group of patterns; 5,6,7,8,9, and 10 bases in length.

When cross-comparisons were made between the five different upstream positional segments (as shown in the main body of the text) the different lengths of regulatory motifs where grouped together since this was the same (and therefore standardized) across the upstream datasets.

FIVE BASE MOTIFS

	upstream1	*upstream2*	*upstream3*	*upstream4*	*upstream5*
Mean	1.107	1.138	1.148	1.156	1.148
Standard Error	0.276	0.348	0.358	0.360	0.357
Median	0.771	0.676	0.688	0.699	0.691
Standard Deviation	0.828	1.045	1.075	1.080	1.070
Skewness	1.344	1.284	1.269	1.235	1.255
Count	9	9	9	9	9
Confidence Level(95.0%)	0.637	0.803	0.827	0.830	0.823

SIX BASE MOTIFS

	upstream1	*upstream2*	*upstream3*	*upstream4*	*upstream5*
Mean	1.408	1.227	1.183	1.183	1.180
Standard Error	0.197	0.178	0.178	0.179	0.177
Median	0.889	0.903	0.907	0.952	0.929
Standard Deviation	1.059	0.957	0.956	0.966	0.955
Skewness	0.825	0.618	0.666	0.709	0.651
Count	29	29	29	29	29
Confidence Level(95.0%)	0.403	0.364	0.364	0.367	0.363

SEVEN BASE MOTIFS

	upstream1	*upstream2*	*upstream3*	*upstream4*	*upstream5*
Mean	1.242	0.997	0.993	1.002	0.968
Standard Error	0.314	0.267	0.285	0.282	0.271
Median	1.007	0.802	0.781	0.833	0.801
Standard Deviation	1.254	1.070	1.139	1.129	1.084
Skewness	3.307	3.294	3.321	3.214	3.189
Count	16	16	16	16	16
Confidence Level(95.0%)	0.668	0.570	0.607	0.601	0.577

EIGHT BASE MOTIFS

	upstream1	*upstream2*	*upstream3*	*upstream4*	*upstream5*
Mean	2.166	1.664	1.569	1.522	1.547
Standard Error	0.366	0.284	0.277	0.270	0.271
Median	1.119	0.783	0.545	0.664	0.594
Standard Deviation	2.808	2.184	2.125	2.074	2.085
Skewness	2.705	2.012	2.147	2.265	2.162
Count	59	59	59	59	59
Confidence Level(95.0%)	0.732	0.569	0.554	0.541	0.543

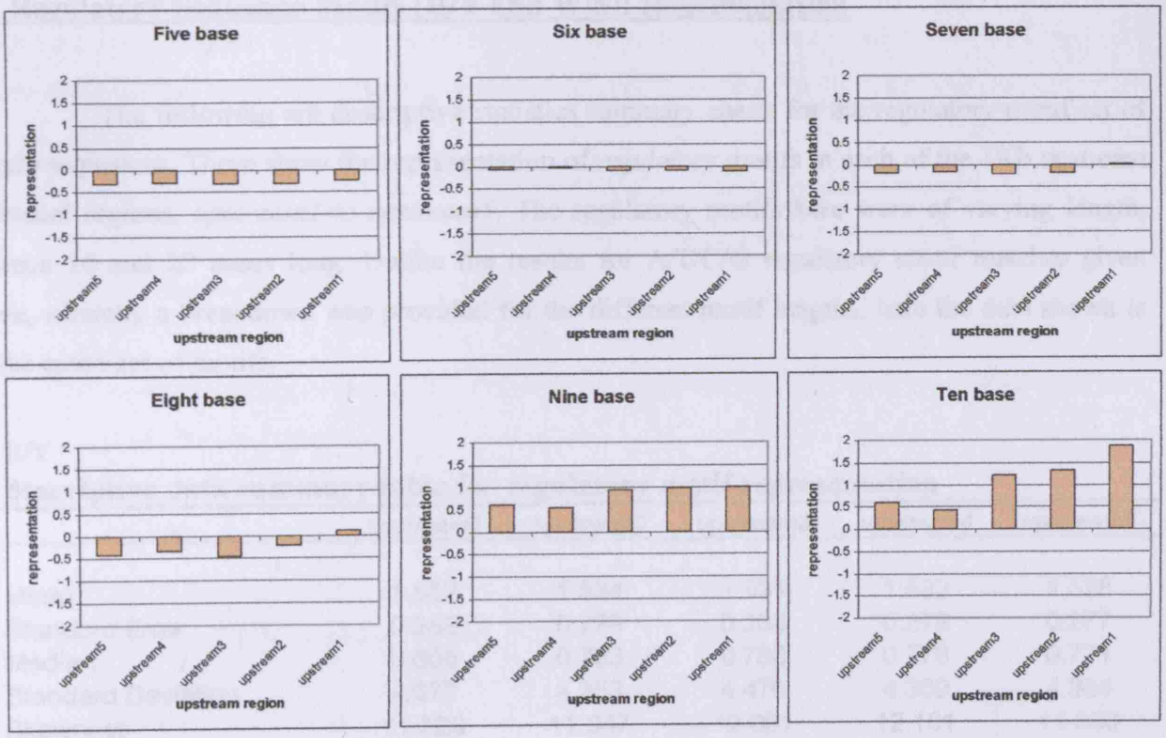
NINE BASE MOTIFS

	upstream1	*upstream2*	*upstream3*	*upstream4*	*upstream5*
Mean	7.143	6.310	5.754	5.650	5.492
Standard Error	2.392	2.050	1.991	1.973	1.909
Median	2.002	1.968	1.950	1.541	1.608
Standard Deviation	9.864	8.454	8.207	8.135	7.870
Skewness	2.380	1.478	1.893	1.869	1.928
Count	17	17	17	17	17
Confidence Level(95.0%)	5.072	4.346	4.220	4.183	4.046

TEN BASE MOTIFS

	upstream1	*upstream2*	*upstream3*	*upstream4*	*upstream5*
Mean	13.756	9.866	9.002	8.433	8.761
Standard Error	4.752	2.735	2.432	2.458	2.477
Median	2.866	2.323	2.207	1.440	1.598
Standard Deviation	22.788	13.115	11.665	11.787	11.881
Skewness	2.291	1.296	1.339	1.519	1.397
Count	23	23	23	23	23
Confidence Level(95.0%)	9.854	5.671	5.044	5.097	5.138

Representation of Regulatory Motif Matches Across the Upstream Positional Regions



Representation of Regulatory Motif Matches Across the Upstream Positional regions (*upstream5*-to-*upstream1*)						
Frequency: median values for each set of regulatory motifs						
Motif length / **Upstream region**	**ten**	**nine**	**eight**	**seven**	**six**	**five**
upstream5	0.598	0.608	-0.406	-0.199	-0.071	-0.309
upstream4	0.440	0.541	-0.336	-0.167	-0.048	-0.301
upstream3	1.207	0.950	-0.455	-0.219	-0.093	-0.312
upstream2	1.323	0.968	-0.217	-0.198	-0.097	-0.324
upstream1	1.866	1.002	0.119	0.007	-0.111	-0.229

The plots and data-table above show the representation of regulatory motifs across the upstream datasets. Each plot contains the number of matches of regulatory motifs of specified length (in bases) within each upstream dataset. The median number of matches is given across the dataset of motifs.

For the majority of the motif lengths, except for the six base motifs, the representation increases on average from *upstream5*-to-*upstream1* in the downstream direction. This effect is more pronounced with the larger motif lengths, i.e. nine and ten base motifs.

D.4 Regulatory Sequence Motifs (R/Y and W/S): Representation

The following are descriptive statistics summary charts for the regulatory motif set of sample sequences. These show the representation of regulatory motifs in each of the 1Kb upstream positional regions; *upstream1*-to-*upstream5*. The regulatory motifs here were of varying length, between 10 and 20 bases long. Unlike the results for A/T/C/G regulatory motif matches given above, whereby a breakdown was provided for the different motif lengths, here the data shown is for the entire set of motifs.

R/Y

descriptive data summary table for regulatory motif representation

	upstream1	*upstream2*	*upstream3*	*upstream4*	*upstream5*
Mean	1.552	1.534	1.531	1.522	1.528
Standard Error	0.258	0.276	0.284	0.278	0.277
Median	0.805	0.783	0.783	0.776	0.771
Standard Deviation	4.077	4.353	4.476	4.389	4.364
Skewness	11.529	11.947	12.031	12.101	11.980
Count	286	286	286	286	286

Confidence Level(95.0%)	0.509	0.543	0.559	0.548	0.545

W/S

descriptive data summary table for regulatory motif representation

	upstream1	*upstream2*	*upstream3*	*upstream4*	*upstream5*
Mean	2.058	2.033	1.922	1.857	1.962
Standard Error	0.340	0.338	0.303	0.271	0.307
Median	0.899	0.994	0.928	0.886	0.926
Standard Deviation	5.088	5.066	4.528	4.056	4.592
Skewness	8.466	8.680	8.400	8.037	8.230
Count	224	224	224	224	224
Confidence Level(95.0%)	0.670	0.667	0.596	0.534	0.605

D.5 Regulatory Motifs Matches in the upstream; Real verses Random

For each of the regulatory motifs its real frequency in the upstream sequence was compared with its theoretical expected frequency for a random sequence of equivalent nucleotide composition. This comparison was made in order to see if there was a significant difference between the real and random sequences.

Ttests were carried to compare the real frequencies of matches within each individual upstream dataset with the expected value for the random sequence. The Ttests were two-tailed with no assumption made regarding equal variance and carried out at the 5% level of significance.

Null hypothesis, Ho: The frequency of regulatory motif matches within a given upstream region is from the same underlying distribution as the frequency of matches in an equivalent random sequence.

Alternative Hypothesis, H_1: The real and random distributions of regulatory sequence matches are different.

The following data-tables show the Ttest results for the motif matches in real and random upstream regions; *upstream1*-to-*upstream5*. Three separate tables are presented for the A/TC/G, R/Y and W/S sequence results.

A/T/C/G datasets

Comparing Real and Random Distributions of regulatory motif matches
TTEST: Paired, Two-tailed at 5% level of significance

	upstream1	*upstream2*	*upstream3*	*upstream4*	*upstream5*
TTEST	0.023	0.300	0.365	0.363	0.383
Ho	reject	accept	accept	accept	accept
Distribution	non-random	random	random	random	random

R/Y datasets

Comparing Real and Random Distributions of regulatory motif matches
TTEST: Paired, Two-tailed at 5% level of significance

	upstream1	*upstream2*	*upstream3*	*upstream4*	*upstream5*
TTEST	0.010	0.013	0.014	0.014	0.013
Ho	reject	reject	reject	reject	reject
Distribution	non-random	non-random	non-random	non-random	non-random

W/S datasets

Comparing Real and Random Distributions of regulatory motif matches
TTEST: Paired, Two-tailed at 5% level of significance

	upstream1	*upstream2*	*upstream3*	*upstream4*	*upstream5*
TTEST	0.008	0.129	0.500	0670	0.603
Ho	reject	accept	accept	accept	accept
Distribution	non-random	random	random	random	random

D.6 Regulatory Motifs Matches in the *whole genome*; Real verses Random

For each of the regulatory motifs its real frequency in the *whole* genome was compared with its theoretical expected frequency for a random sequence of equivalent nucleotide composition. This was carried out and the results (in the table below) are presented chromosome-by-chromosome. The comparison was made in order to see if there was a significant difference between the real and random sequences.

Ttests were carried to compare the real frequencies of matches within each individual chromosome with the expected value for the random sequence. The Ttests were two-tailed with no assumption made regarding equal variance and carried out at the 5% level of significance.

Null hypothesis, Ho: The frequency of regulatory motif matches within the chromosome is from the same underlying distribution as the frequency of matches in an equivalent chromosome with a randomized sequence.

Alternative Hypothesis, H_1: The real and random sequence distributions for regulatory motif matches are different.

The following data-tables show the Ttest results for the motif matches in real and random chromosomes. Three separate columns are presented for the A/TC/G, R/Y and W/S sequence results. Each column shows the Ttest statistic.

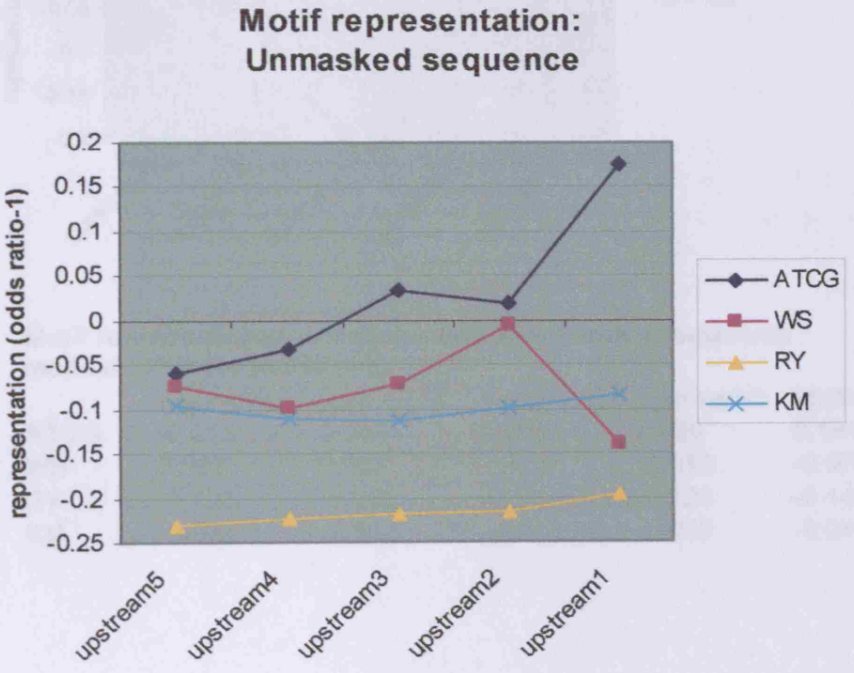
For the A/T/C/G sequences the results of the Ttest reveal that at the 5% level of significance the null hypothesis can be accepted for all of the twenty chromosomes. Therefore the regulatory sequence matches occur at the level that is expected in the random sequence, i.e. at the random level. The W/S sequence results are the same, in that the motif matches occurring at the random level in all the chromosomes. The R/Y results on the other hand are very different. Here the null hypothesis can be rejected at the 5% level of significance and it can be concluded that for all chromosomes the real and randomized sequence distributions are significantly different.

Comparing Real and Random Distributions of regulatory motif matches
TTEST: Paired, Two-tailed at 5% level of significance

chromosome	A/T/C/G -Ttest	R/Y -Ttest	W/S -Ttest
1	0.360	0.008	0.885
2	0.355	0.007	0.671
3	0.379	0.007	0.428
4	0.369	0.006	0.561
5	0.372	0.006	0.530
6	0.351	0.007	0.610
7	0.363	0.008	0.997
8	0.356	0.007	0.575
9	0.360	0.007	0.899
10	0.361	0.007	0.657
11	0.343	0.007	0.876
12	0.359	0.008	0.786
13	0.361	0.006	0.697
14	0.360	0.008	0.947
15	0.373	0.008	0.768
16	0.356	0.010	0.954
17	0.348	0.011	0.664
18	0.345	0.006	0.586
19	0.333	0.013	0.544
20	0.330	0.009	0.647
21	0.321	0.006	0.652
22	0.343	0.010	0.890
X	0.378	0.006	0.375
Y	0.418	0.007	0.074
Ho Distribution	**accept random**	**reject non-random**	**accept random**

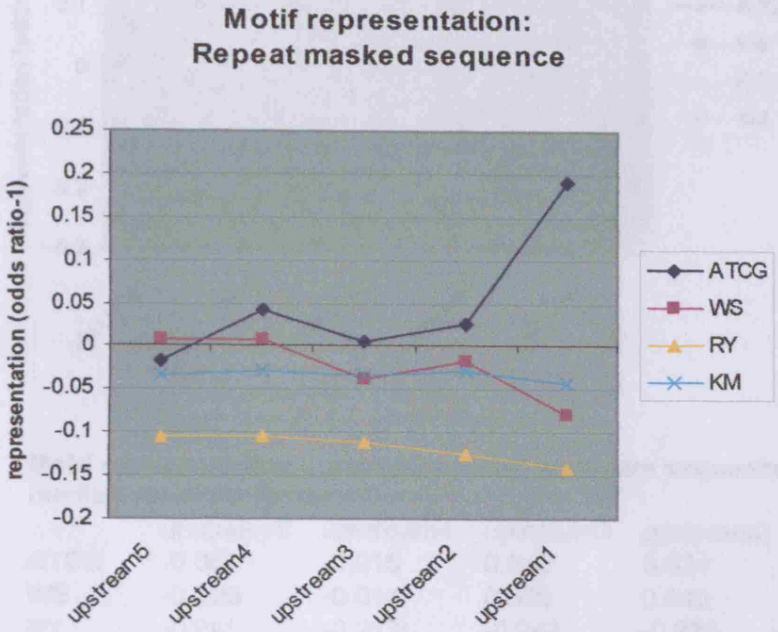
D.7 Motif representation on the sense strand in repeat-free sequences verses repeat-containing sequences

The following results show that the relative motif representation trends across the 5Kb upstream sequence generally remains unchanged whether the sequence is masked for repeats or is left unmasked. The main difference between repeat masked and unmasked datasets is that in the masked datasets all the motif representation values are closer to the random (zero) value. The one notable exception is *upstream1* for which the motif representation is more distant from randomness in the repeat-free dataset than the equivalent repeat-containing dataset. Also in the repeat masked datasets there is a more pronounced change (for ATCG and W/S sequences) between *upstream1* and the further upstream datasets. Here, between *upstream2*-to-*upstream5* the motif representation appears at the random level throughout.



Motif representation in unmasked upstream sequence: median value for (odds ratio-1)

	upstream5	upstream4	upstream3	upstream2	upstream1
ATCG	-0.060	-0.034	0.034	0.020	0.176
WS	-0.074	-0.098	-0.072	-0.006	-0.141
RY	-0.229	-0.224	-0.217	-0.217	-0.195
XY	-0.097	-0.111	-0.113	-0.098	-0.085



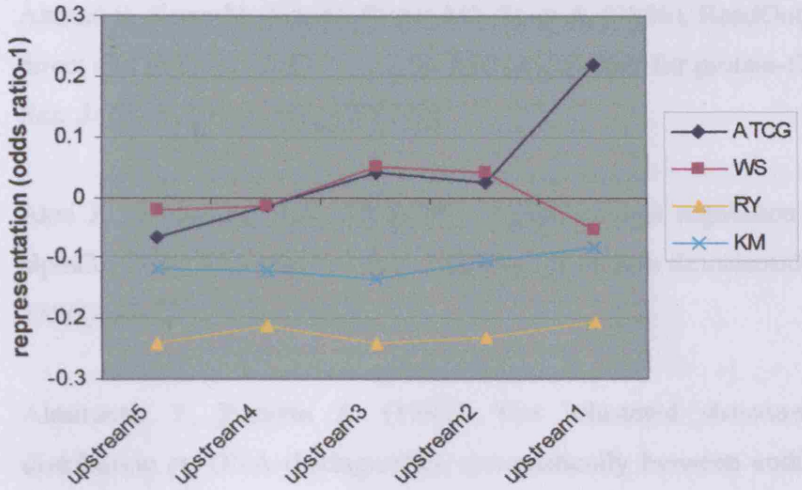
Motif representation in repeat masked upstream sequence: median value for (odds ratio-1)

	upstream5	upstream4	upstream3	upstream2	upstream1
ATCG	-0.018	0.043	0.006	0.026	0.191
WS	0.007	0.007	-0.038	-0.016	-0.078
RY	-0.103	-0.105	-0.112	-0.124	-0.140
KM	-0.034	-0.030	-0.033	-0.029	-0.042

D.8 Motif representation in the anti-sense strand

So far the results shown have been for motif matches at the sense strand. The following result is for the anti-sense strand (unmasked upstream sequence). In general the representation trends across the upstream sequences are similar for the two strands. The range of representation values are the same. One notable difference is in the W/S *upstream1* region. For this dataset the representation is closer to the random value in the non-transcribed strand than in the transcribed strand. This result suggests some asymmetry in the presence of W/S content of motifs between the two strands.

Motif representation:
The anti-sense strand



Motif representation in repeat masked upstream sequence:
median value for (odds ratio-1)

	upstream5	upstream4	upstream3	upstream2	upstream1
ATCG	-0.067	-0.015	0.040	0.024	0.220
WS	-0.020	-0.014	0.050	0.040	-0.054
RY	-0.241	-0.212	-0.243	-0.232	-0.205
KM	-0.118	-0.120	-0.135	-0.106	-0.082

References

Ahmad S, Kono H, Arauzo-Bravo MJ, Sarai A. (2006). ReadOut: structure-based calculation of direct and indirect readout energies and specificities for protein-DNA recognition. *Nucleic Acids Res. 34(Web Server issue):W124-7.*

Akai J, Kimura A, Hata RI. (1999) Transcriptional regulation of the human type I collagen alpha2 (COL1A2) gene by the combination of two dinucleotide repeats. *Gene. 1999 Oct 18; 239(1):65-73.*

Almirantis Y, Provata A. (1997). The "clustered structure" of the purines/pyrimidines distribution in DNA distinguishes systematically between coding and non-coding sequences. *Bull Math Biol. Sep; 59(5):975-92.*

Amano N, Ohfuku Y, Suzuki M. (1997) Genomes and DNA conformation. *Biol Chem. 378(12):1397-404.*

Angelier N, Bonnanfant-Jais ML, Herberts C, Lautredou N, Moreau N, N'Da E, Penrad-Mobayed M, Rodriguez-Martin ML, Sourrouille P. (1990). Chromosomes of amphibian oocytes as a model for gene expression: significance of lampbrush loops. *Int J Dev Biol. 34(1):69-80.*

Aso T, Serizawa H, Conaway RC, Conaway JW. (1994) A TATA sequence-dependent transcriptional repressor activity associated with mammalian transcription factor IIA. EMBO J. Jan 15;13(2):435-45.

Austin RJ, Biggin MD. (1995) A domain of the even-skipped protein represses transcription by preventing TFIID binding to a promoter: repression by cooperative blocking. Mol Cell Biol. Sep;15(9):4683-93.

Baldi P, Chauvin Y, Brunak S, Gorodkin J, Pedersen AG. (1998) Computational applications of DNA structural scales. *Proc Int Conf Intell Syst Mol Biol. 6:35-42.*

Banavali NK, Roux B. (2005) Free energy landscape of A-DNA to B-DNA conversion in aqueous solution. *J Am Chem Soc; 127(18):6866-76.*

Benos PV, Lapedes AS, Stormo GD. (2002) Probabilistic code for DNA recognition by proteins of the EGR family. *J Mol Biol; 323(4):701-27.*

Bhaumik SR, Raha T, Aiello DP, Green MR. (2004) In vivo target of a transcriptional activator revealed by fluorescence resonance energy transfer. Genes Dev. Feb 1;18(3):333-43.

Birney E, Stamatoyannopoulos JA and the ENCODE Project Consortium, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature.* 2007 14;447(7146):782-3.

Blaisdell BE. (1983) A prevalent persistent global nonrandomness that distinguishes coding and non-coding eucaryotic nuclear DNA sequences. *J Mol Evol; 19(2):122-33.*

Blake MC, Jambou RC, Swick AG, Kahn JW, Azizkhan JC. (1990) Transcriptional initiation is controlled by upstream GC-box interactions in a TATAA-less promoter. Mol Cell Biol. Dec;10(12):6632-41.

Brahmachari SK, Meera G, Sarkar PS, Balagurumoorthy P, Tripathi J, Raghavan S, Shaligram U, Pataskar S. (1995). Simple repetitive sequences in the genome: structure and functional significance.Electrophoresis.16(9):1705-14.

Breslauer KJ, Frank R, Blocker H, Marky LA. (1996). Predicting DNA duplex stability from the base sequence. Proc. Natl. Acad. Sci. USA. 83: 3746-50

Brickner AG, Koop BF, Aronow BJ, Wiginton DA. (1999) Genomic sequence comparison of the human and mouse adenosine deaminase gene regions. Mamm Genome. Feb;10(2):95-101.

Brivanlou AH, Darnell JE Jr. (2002). Signal transduction and the control of gene expression. *Science; 295(5556):813-8.*

Burge C, Campbell AM, Karlin S. (1992). Over- and under-representation of short oligonucleotides in DNA sequences. *Proc Natl Acad Sci U S A; 89(4):1358-62.*

Burley SK, Roeder RG. (1996) Biochemistry and structural biology of transcription factor IID (TFIID). Annu Rev Biochem.;65:769-99.

Butler JE, Kadonaga JT. (2001) Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. Genes Dev. Oct 1;15(19):2515-9.

Calladine CR, Drew HR. (1986). Principles of sequence-dependent flexure of DNA. *J Mol Biol; 192(4):907-18.*

Calladine CR, Drew HR., Luisi BF, Travers AA. (2004). Understanding DNA; the molecule and how it works. *Third edition, Elsevier Academic press.*

Campbell A, Mrazek J, Karlin S. (1999) Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc Natl Acad Sci U S A; 96(16):9184-9.*

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet; 22(3):231-8. Erratum in: Nat Genet 1999 Nov;23(3):373.*

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engström PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet. Jun;38(6):626-35. Epub 2006 Apr 28. Erratum in: Nat Genet. 2007
Sep;39(9):1174.

Celniker SE, Drewell RA. (2007). Chromatin looping mediates boundary element promoter interactions. *Bioessays; 29(1):7-10.*

Chen J, Kinyamu HK, Archer TK. (2006). Changes in attitude, changes in latitude: nuclear receptors remodeling chromatin to regulate transcription. *Mol Endocrinol; 20(1):1-13.*

Claverie JM. (2001). Gene number: What if there are only 30,000 human genes? *Science; 291(5507):1255-7.*

Conaway RC, Conaway JW. (. 1997) General transcription factors for RNA polymerase II. Prog Nucleic Acid Res Mol Biol;56:327-46.

Conaway RC, Sato S, Tomomori-Sato C, Yao T, Conaway JW. (2005) The mammalian Mediator complex and its role in transcriptional regulation. Trends Biochem Sci. May;30(5):250-5.

Cooper SJ, Trinklein ND, Anton ED, Nguyen L, Myers RM. (2006). Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res; 16(1):1-10.*

Courey AJ, Jia S. (2001) Transcriptional repression: the long and the short of it. Genes Dev. Nov 1;15(21):2786-96.

Dilworth FJ, Chambon P. (2001). Nuclear receptors coordinate the activities of chromatin remodeling complexes and coactivators to facilitate initiation of transcription. *Oncogene; 20(24):3047-54.*

Donaill DA. (2000). A parity code interpretation of nucleotide alphabet composition. *Chem Commun (Camb); (18):2062-3.*

Donaldson IJ, Gottgens B. (2007). CoMoDis: composite motif discovery in mammalian genomes. *Nucleic Acids Res; 35(1):e1. Epub 2006 Nov 27.*

Edmondson DG, Roth SY. (1996). Chromatin and transcription. *FASEB J; 10(10):1173-82.*

El Hassan MA, Calladine CR. (1998). Two distinct modes of protein-induced bending in DNA. *J Mol Biol; 282(2):331-43.*

El Hassan MA, Calladine CR. (1996). Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. *J Mol Biol; 259(1):95-103.*

El Hassan MA, Calladine CR. (1995). The assessment of the geometry of dinucleotide steps in double-helical DNA; a new local calculation scheme. *J Mol Biol; 251(5):648-64.*

Elrod-Erickson M, Benson TE, Pabo CO. (1998). High-resolution structures of variant Zif268-DNA complexes: implications for understanding zinc finger-DNA recognition. *Structure; 6(4):451-64.*

Eriksson MA, Nilsson L. (1995). Structure, thermodynamics and cooperativity of the glucocorticoid receptor DNA-binding domain in complex with different response elements.

Molecular dynamics simulation and free energy perturbation studies. *J Mol Biol; 253(3):453-72.*

Feuerstein BG, Pattabiraman N, Marton LJ. (1990). Molecular mechanics of the interactions of spermine with DNA: DNA bending as a result of ligand binding. *Nucleic Acids Res; 18(5):1271-82.*

Fukue Y, Sumida N, Tanase J, Ohyama T. (2005). A highly distinctive mechanical property found in the majority of human promoters and its transcriptional relevance. *Nucleic Acids Res; 33(12):3821-7.*

Fukue Y, Sumida N, Nishikawa J, Ohyama T. (2004). Core promoter elements of eukaryotic genes have a highly distinctive mechanical property. *Nucleic Acids Res; 32(19):5834-40.*

Gao J, Qi Y, Cao Y, Tung WW. (2005). Protein coding sequence identification by simultaneously characterizing the periodic and random features of DNA sequences. *J Biomed Biotechnol; 2005(2):139-46.*

Garvie CW, Wolberger C. (2001). Recognition of specific DNA sequences. *Mol Cell; 8(5):937-46.*

Gaudreau L, Schmid A, Blaschke D, Ptashne M, Hörz W.
RNA polymerase II holoenzyme recruitment is sufficient to remodel chromatin at the yeast PHO5 promoter.
Cell. 1997 Apr 4;89(1):55-62.

Gentles AJ, Karlin S. (2001). Genome-Scale Compositional Comparisons in Eukaryotes. *11:540-546*

Gray S, Levine M. (1996) Short-range transcriptional repressors mediate both quenching and direct repression within complex loci in Drosophila. Genes Dev. Mar 15;10(6):700-10.

Grosschedl R, Giese K, Pagel J. (1994) HMG domain proteins: architectural elements in the assembly of nucleoprotein structures. Trends Genet. Mar;10(3):94-100.

Guo Y, Jamison DC. (2005). The distribution of SNPs in human gene regulatory regions. *BMC Genomics; 6:140.*

Hampsey M. (1998) Molecular genetics of the RNA polymerase II general transcriptional machinery. Microbiol Mol Biol Rev. Jun;62(2):465-503.

Hardison, R.C. (2000). Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* 2000;**16**:369–372.

Hoglund A, Kohlbacher O. (2004). From sequence to structure and back again: approaches for predicting protein-DNA binding. *Proteome Sci; 2(1):3.*

Huang F, Cabaud O, Verthuy C, Hueber AO, Ferrier P. (2003). Alpha beta T-cell development is not affected by inversion of TCR beta gene enhancer sequences: polar enhancement of gene expression regardless of enhancer orientation. *Immunology; 109(4):510-4.*

Hunter CA. (1993). Sequence-dependent DNA structure. The role of base stacking interactions. *J Mol Biol;230(3):1025-54.*

Iglesias AR, Kindlund E, Tammi M, Wadelius C. (2004). Some microsatellites may act as novel polymorphic cis-regulatory elements through transcription factor binding. *Gene; 341: 149-65.*

Iyer V, Struhl K. (1995). Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J; 14(11):2570-9.*

Jamieson AC, Kim SH, Wells JA. (1994). In vitro selection of zinc fingers with altered DNA-binding specificity. *Biochemistry; 33(19):5689-95.*

Jernigan RW, Baran RH. (2002). Pervasive properties of the genomic signature. *BMC Genomics; 3(1):23.*

Kadonaga JT. (2002) The DPE, a core promoter element for transcription by RNA polymerase II. Exp Mol Med Sep 30;34(4):259-64.

Karlin S, Brendel V. (1993). Patchiness and correlations in DNA sequences. *Science; 259(5095):677-80.*

Karlin S, Burge C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet;11(7):283-90..*

Karlin S, Campbell AM, Mrazek J. (1998). Comparative DNA analysis across diverse genomes. *Annu Rev Genet; 32:185-225.*

Karlin S, Ladunga I. (1994). Comparisons of eukaryotic genomic sequences. *Proc Natl Acad Sci U S A; 91(26):12832-6.*

Karlin S, Mrazek J. (1996). What drives codon choices in human genes? *J Mol Biol; 262(4):459-72.*

Karlin S, Mrazek J. (1997). Compositional differences within and between eukaryotic genomes. *Proc Natl Acad Sci U S A; 94(19):10227-32.*

Kaufmann J, Smale ST. (1994) Direct recognition of initiator elements by a component of the transcription factor IID complex. Genes Dev. Apr 1;8(7):821-9.

Kel OV, Romaschenko AG, Kel AE, Wingender E, Kolchanov NA. (1995). A compilation of composite regulatory elements affecting gene transcription in vertebrates. *Nucleic Acids Res; 23(20):4097-103.*

Kel-Margoulis OV, Romashchenko AG, Kolchanov NA, Wingender E, Kel AE. (2000) COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation. *Nucleic Acids Res; 28(1):311-5.*

Kim Y, Geiger JH, Hahn S, Sigler PB. (1993) Crystal structure of a yeast TBP/TATA-box complex. Nature. Oct 7;365(6446):512-20.

Kornberg RD.
The molecular basis of eukaryotic transcription.
Proc Natl Acad Sci U S A. 2007 Aug 7;104(32):12955-61.

Koudelka GB, Harrison SC, Ptashne M. (1987). Effect of non-contacted bases on the affinity of 434 operator for 434 repressor and Cro. *Nature; 326(6116):886-8.*

Lamoureux JS, Maynes JT, Glover JN. (2004). Recognition of 5'-YpG-3' sequences by coupled stacking/hydrogen bonding interactions with amino acid residues. *J Mol Biol; 335(2):399-408.*

Larhammar D, Chatzidimitriou-Dreismann CA. (1993). Biological origins of long-range correlations and compositional variations in DNA.

Nucleic Acids Res. 1993 Nov 11; 21(22):5167-70.

Lee CK, Shibata Y, Rao B, Strahl BD, Lieb JD. (2004) Evidence for nucleosome depletion at active regulatory regions genome-wide. Nat Genet, Aug;36(8):900-5. Epub 2004 Jul 11.

Lewis JD, Meehan RR, Henzel WJ, Maurer-Fogy I, Jeppesen P, Klein F, Bird A. (1992) Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. Cell. Jun 12;69(6):905-14.

Lewis BA, Reinberg D. (2003) The mediator coactivator complex: functional and physical roles in transcriptional regulation. J Cell Sci. Sep 15;116(Pt 18):3667-75.

Lim CY, Santoso B, Boulay T, Dong E, Ohler U, Kadonaga JT. (2004) The MTE, a new core promoter element for transcription by RNA polymerase II. Genes Dev. Jul 1;18(13):1606-17.

Li H, Ilin S, Wang W, Duncan EM, Wysocka J, Allis CD, Patel DJ. (2006) Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. Nature. Jul 6;442(7098):91-5. Epub 2006 May 21.

Li XY, Virbasius A, Zhu X, Green MR. (1999) Enhancement of TBP binding by activators and general transcription factors. Nature. Jun 10;399(6736):605-9.

Li W, Miramontes P. (2006). Large-scale oscillation of structure-related DNA sequence features in human chromosome 21. *Phys Rev E Stat Nonlin Soft Matter Phys; 74(2 Pt 1):021912.*

Lobry JR. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol;13(5):660-5.

Long F, Liu H, Hahn C, Sumazin P, Zhang MQ, Zilberstein A. (2004). Genome-wide prediction and analysis of function-specific transcription factor binding sites. *In Silico Biol; 4(4):395-410.*

Louie E, Ott J, Majewski J. (2003). Nucleotide frequency variation across human genes. *Genome Res; 13(12):2594-601.*

Luscombe NM, Laskowski RA, Thornton JM. (2001). Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res; 29(13):2860-74.*

Luscombe NM, Thornton JM. (2002). Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J Mol Biol; 320(5):991-1009.*

Mandel-Gutfreund Y, Schueler O, Margalit H. (1995). Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *J Mol Biol; 253(2):370-82.*

Majewski J, Ott J. (2002). Distribution and characterization of regulatory elements in the human genome. *Genome Res; 12(12):1827-36.*

Merika M, Thanos D. (2001) Enhanceosomes. Curr Opin Genet Dev. Apr;11(2):205-8.

Mrazek J, Karlin S. (1998). Strand compositional asymmetry in bacterial and large viral genomes. Proc Natl Acad Sci;95(7):3720-5.

Mrazek J, Kypr J. (1992). DNABIND: an interactive microcomputer program searching for nucleotide sequences that may code for conserved DNA-binding protein motifs. *Comput Appl Biosci; 8(4):401-4.*

Ng HL, Kopka ML, Dickerson RE. (2000) The structure of a stable intermediate in the A <--> B DNA helix transition. *Proc Natl Acad Sci U S A; 97(5):2035-9.*

Niu DK, Lin K, Zhang DY. (2003). Strand compositional asymmetries of nuclear DNA in eukaryotes. J Mol Evol; 57(3):325-34.

Pabo CO, Nekludova L. (2000). Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? *J Mol Biol; 301(3):597-624.*

Packer MJ, Dauncey MP, Hunter CA. (2000). Sequence-dependent DNA structure: dinucleotide conformational maps. *J Mol Biol; 295(1):71-83.*

Panne D. (2008) The enhanceosome. Curr Opin Struct Biol. Apr;18(2):236-42.

Peng CK, Buldyrev SV, Goldberger AL, Havlin S, Mantegna RN, Simons M, Stanley HE. (1995). Statistical properties of DNA sequences. *Physica A; 221:180-92.*

Ptashne M, Gann A. (1997) Transcriptional activation by recruitment. Nature. Apr 10;386(6625):569-77.

Rajendrakumar P, Biswal AK, Balachandran SM, Srinivasarao K, Sundaram RM. (2007). Simple sequence repeats in organellar genomes of rice: frequency and distribution in genic and intergenic regions. *Bioinformatics; 23(1):1-4. Epub 2006 Oct 31.*

Reese JC. (2003). Basal transcription factors. *Curr Opin Genet Dev. 2003 Apr; 13(2):114-8.*

Reich Z, Friedman P, Scolnik Y, Sussman JL, Minsky A. (1993). On the metastability of left-handed DNA motifs. *Biochemistry; 32(8):2116-9.*

Rigoutsos I, Floratos A. (1998). Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics;14(1):55-67. Erratum in: Bioinformatics 1998; 14(2):229.*

Ringrose L, Chabanis S, Angrand PO, Woodroofe C, Stewart AF. (1999). Quantitative comparison of DNA looping in vitro and in vivo: chromatin increases effective DNA flexibility at short distances. *EMBO J; 18(23):6630-41.*

Rippe K, von Hippel PH, Langowski J. (1995). Action at a distance: DNA-looping and initiation of transcription. *Trends Biochem Sci; 20(12):500-6.*

Roytberg MA. (1992). A search for common patterns in many sequences. *Comput Appl Biosci; 8(1):57-64.*

Rozenberg H, Rabinovich D, Frolow F, Hegde RS, Shakked Z. (1998). Structural code for DNA recognition revealed in crystal structures of papillomavirus E2-DNA targets. *Proc Natl Acad Sci U S A; 95(26):15194-9.*

Saxonov S, Berg P, Brutlag DL. (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A. Jan 31;103(5):1412-7.

Schatz T, Langowski J. Curvature and sequence analysis of eukaryotic promoters. *J Biomol Struct Dyn; 15(2):265-75.*

Seeman NC, Rosenberg JM, Rich A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proc Natl Acad Sci U S A; 73(3):804-8.*

Shapiro JA, von Sternberg R. (2005) Why repetitive DNA is essential to genome function. Biol Rev Camb Philos Soc;80(2):227-50.

Sheridan SD, Opel ML, Hatfield GW. (2001). Activation and repression of transcription initiation by a distant DNA structural transition. *Mol Microbiol; 40(3):684-90.*

Shioiri C, Takahata N. (2001). Skew of mononucleotide frequencies, relative abundance of dinucleotides, and DNA strand asymmetry. J Mol Evol;53(4-5):364-76.
Shomer B, Yagil G, (1999). Long W tracts are over-represented in the *E.Coli* and *H. Influenzae* genomes. Nucleic Acids Res.; 27; 4480-4491

Singal R, Ferris R, Little JA, Wang SZ, Ginder GD. (1997) Methylation of the minimal promoter of an embryonic globin gene silences transcription in primary erythroid cells. Proc Natl Acad Sci U S A. Dec 9;94(25):13724-9.

Sipos L, Gyurkovics H. (2005). Long-distance interactions between enhancers and promoters. *FEBS J; 272(13):3253-9.*

Smale ST, Schmidt MC, Berk AJ, Baltimore D. 1990) Transcriptional activation by Sp1 as directed through TATA or initiator: specific requirement for mammalian transcription factor IID.
Proc Natl Acad Sci U S A. Jun;87(12):4509-13.

Smith HO, Annau TM, Chandrasegaran S. (1990). Finding sequence motifs in groups of functionally related proteins. *Proc Natl Acad Sci U S A; 87(2):826-30.*

Stepanova M, Tiazhelova T, Skoblov M, Baranova A. (2005). A comparative analysis of relative occurrence of transcription factor binding sites in vertebrate genomes and gene promoter areas. *Bioinformatics; 21(9):1789-96.*

Stormo GD. (2000). DNA binding sites: representation and discovery. *Bioinformatics; 16(1):16-23.*

Subramanian S, Kumar S. (2003). Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. *Genome Res; 13(5):838-44.*

Suyama M, Nishioka T, Oda J. (1995). Searching for common sequence patterns among distantly related proteins. *Protein Eng; 8(11):1075-80.*

Suzuki M, Yagi N. (1994). DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families. *Proc Natl Acad Sci U S A; 91(26):12357-61.*

Suzuki Y, Yamashita R, Shirota M, Sakakibara Y, Chiba J, Mizushima-Sugano J, Nakai K, Sugano S. (2004). Sequence comparison of human and mouse genes reveals a homologous block structure in the promoter regions. *Genome Res; 14(9):1711-8.*

Suzuki Y, Yamashita R, Nakai K, Sugano S. (2002). DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res; 30(1):328-31.*

Sved J, Bird A. (1990). The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc Natl Acad Sci U S A; 87(12):4692-6.*

Teng Y, Waters R. (2000). Excision repair at the level of the nucleotide in the upstream control region, the coding sequence and in the region where transcription terminates of the Saccharomyces cerevisiae MFA2 gene and the role of RAD26. Nucleic Acids Res; 1;28(5):1114-9.

Thurman RE, Day N, Noble WS, Stamatoyannopoulos JA. (2007) Identification of higher-order functional domains in the human ENCODE regions. Genome Res. Jun;17(6):917-27.

Tiesman J, Rizzino A. (1990). Nucleotide sequence of the 5'-flanking region of the mouse k-FGF oncogene exhibits an alternating purine:pyrimidine motif with the potential to form Z-DNA. *Gene; 96(2):311-2.*

Tran P, Leclerc D, Chan M, Pai A, Hiou-Tim F, Wu Q, Goyette P, Artigas C, Milos R, Rozen R. (2002). Multiple transcription start sites and alternative splicing in the methylenetetrahydrofolate reductase gene result in two enzyme isoforms. *Mamm Genome; 13(9):483-92.*

Tripathi J, Brahmachari SK. (1991). Distribution of simple repetitive (TG/CA)n and (CT/AG)n sequences in human and rodent genomes. *J Biomol Struct Dyn; 9(2):387-97.*

Tsai FT, Sigler PB. (2000). Structural basis of preinitiation complex assembly on human pol II promoters. EMBO J; 19(1):25-36.

Venter JC, ...Zhu X. (2001) The sequence of the human genome. *Science; 291(5507):1304-51.*

Wang AH, Gessner RV, van der Marel GA, van Boom JH, Rich A. (1985). Crystal structure of Z-DNA without an alternating purine-pyrimidine sequence. *Proc Natl Acad Sci U S A; 82(11):3611-5.*

Werner F, Weinzierl RO. (2005). Direct modulation of RNA polymerase core functions by basal transcription factors. *Mol Cell Bio; 25(18):8344-55.*

Wingender E, Chen X, Fricke E, Geffers R, Hehl R, Liebich I, Krull M, Matys V, Michael H, Ohnhauser R, Pruss M, Schacherer F, Thiele S, Urbach S. (2001). The TRANSFAC system on gene expression regulation. *Nucleic Acids Res; 29(1):281-3.*

Wu SY, Zhou T, Chiang CM. (2006) Human mediator enhances activator-facilitated recruitment of RNA polymerase II and promoter recognition by TATA-binding protein (TBP) independently of TBP-associated factors. *Mol Cell Biol; 23(17):6229-42.*

Xue W, Wang J, Shen Z, Zhu H. (2004). Enrichment of transcriptional regulatory sites in non-coding genomic region. *Bioinformatics. 2004 Mar 1;20(4):569-75.*

Yagil G. (2006). DNA tracts composed of only two bases concentrate in promoters. Genomics 87; 591-597

Yamashita R, Suzuki Y, Wakaguri H, Tsuritani K, Nakai K, Sugano S. (2006). DBTSS: DataBase of Human Transcription Start Sites, progress report 2006. *Nucleic Acids Res; 34(Database issue):D86-9.*

Zhang ZD, Paccanaro A, Fu Y, Weissman S, Weng Z, Chang J, Snyder M, Gerstein MB. (2007) Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions. Genome Res. 17(6):787-97.

Zhao F, Xuan Z, Liu L, Zhang MQ. (2005). TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies. *Nucleic Acids Res;33(Database issue):D103-7.*

Zhurkin VB. (1983). Specific alignment of nucleosomes on DNA correlates with periodic distribution of purine-pyrimidine and pyrimidine-purine dimers. *FEBS Lett; 158(2):293-7.*